COMP 7070 Advanced Topics in AI and ML April 2, 2025

Lecture 11: Duality and Flow model

Instructor: Yifan Chen *Scribes: Jiaer Xia, Haoliang Han* *Proof reader: Xiong Peng*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

11.1 Duality in Optimization

11.1.1 General duality concept

For a set A and a point x, the duality concept can be expressed as follows:

$$d(A,x) = \min_{a \in A} d(x,a) = \max_{s \text{ separates } A,x} d(x,s).$$

where s is the boundary separating point x from set A.

11.1.2 Dual form derivation

Consider the primal optimization problem:

$$\min_{x \in D} f(x), \text{ subject to some constraints.}$$

Here, $f(x)$ is a function to be minimized over the domain D. Transforming it into the Lagrangian form with the use of Lagrange multipliers λ and y for the inequality constraints $g(x) \le 0$ and the equality constraint $h(x) = 0$, respectively, we obtain the Lagrangian:

$$L(x,\lambda,y) = f(x) + \lambda^\top g(x) + y^\top h(x).$$

The dual problem involves flipping the min and max, expressed as:

$$\max_{\lambda \ge 0, y} \min_{x \in D} L(x,\lambda,y).$$

This forms the dual optimization problem, where λ is the vector of Lagrange multipliers for the inequality constraints, and $g(x) \ge 0$ ensures feasibility.

11.2 Wasserstein Case

Considering the Wasserstein distance, a special case in optimal transport:

$$\inf_{\pi \in \Pi} \int_{X \times Y} c(x,y) \cdot \pi(x,y) dx dy,$$

where $\pi \in \Pi\left(P_X, P_Y\right)$ is a transport plan within the set of all possible plans between the probability measures P_X and P_Y, and $c(x,y)$ is the cost function, often chosen to be a distance measure between x and y.

Consider any transport map $\pi(x,y)$ without constraints. The optimality condition can be given as:

$$\mu(x) - \int_Y \pi(x,y) dy = 0.$$

Additionally, for π in some set of transport plans UV, we have the following conditions:

$$\begin{aligned}\nu(y) - \textstyle\int_X \pi(x,y)dx = 0,\\ -\pi(x,y) \leq 0.\end{aligned}$$

The Lagrangian $L(\Pi, \phi, \psi, \tau)$ can be expressed as:

$$\begin{aligned}L(\Pi, \phi, \psi, \tau) =& C_k(\pi) + \int \phi(x)\left[\mu(x) - \int_Y \pi(x,y)dy\right]dx\\ &+ \int \psi(y)\left[v(y) - \int_X \pi(x,y)dx\right]dy\\ &+ \iint \tau(x,y)[-\pi(x,y)]dxdy.\end{aligned}$$

The primal problem can be stated as finding the infimum of the Lagrangian form:

$$\inf_{\pi\in\Pi} \sup_{\phi,\psi,\tau\geq 0} L(\Pi, \phi, \psi, \tau).$$

Moreover, we derive the inner integration in the Lagrangian formulation concerning measures and potential functions:

$$\int L(\Pi, \phi, \psi, \tau) = \inf_{\Pi} \int \pi(x,y)[c(x,y) - \tau(x,y) - \phi(x), -\psi(y)]dy.$$

And we have:

$$c(x,y) - \tau(x,y) - \phi(x) - \psi(y) = 0, \quad \text{when } \Pi(x,y) > 0.$$

Upon eliminating terms, we have:

$$c(x,y) - \phi(x) - \psi(y) \geq 0.$$

Thus we have the dual formulation:

$$\sup_{\phi,\psi}\left(\int \phi(x)du(x) + \int \psi(y)dv(y)\right), \text{ subject to } \phi(x) + \psi(y) \leq c(x,y).$$

11.3 Functional derivative and first variation

11.3.1 Functional derivative

A functional is a mapping from a space of functions to the real numbers. For example, if $J(y)$ is a functional of a function $y(x)$, it might be defined as:

$$J(y) = \int_a^b F(x, y(x), y'(x))dx.$$

where F is a given function of x, $y(x)$, and $y'(x)$.

The *functional derivative* of $J(y)$ with respect to $y(x)$ measures how $J(y)$ changes when $y(x)$ is varied. It is denoted by $\frac{\delta J}{\delta y(x)}$.

To find the functional derivative, consider a small perturbation of $y(x)$ given by $y(x)+\epsilon\eta(x)$, where $\eta(x)$ is an arbitrary function and ϵ is a small parameter. The change in $J(y)$ is given by:

$$J(y+\epsilon\eta) - J(y) = \epsilon\int_a^b \left(\frac{\delta J}{\delta y(x)}\right)\eta(x)dx + o(\epsilon).$$

The functional derivative $\frac{\delta J}{\delta y(x)}$ is the function such that this integral accounts for the first-order change in $J(y)$.

11.3.2 First variation

The *first variation* $\delta J(y;\eta)$ is the linear term in the expansion of $J(y+\epsilon\eta)$ around $\epsilon = 0$:

$$\delta J(y;\eta) = \int_a^b \left(\frac{\delta J}{\delta y(x)}\right)\eta(x)dx.$$

If $\delta J(y;\eta) = 0$ for all admissible $\eta(x)$, then $y(x)$ is a stationary point of the functional $J(y)$, often corresponding to an extremum (minimum or maximum).

11.3.3 Euler-Lagrange equation

The Euler-Lagrange equation is a fundamental result in the calculus of variations, providing the conditions for a function to be an extremum of a functional. Consider a functional $J(y)$ of the form:

$$J(y) = \int_a^b L(x, y(x), y'(x))dx.$$

where L is the Lagrangian, a function of x, $y(x)$, and $y'(x)$. We introduce a small perturbation to the function $y(x)$ as $y(x) + \epsilon\eta(x)$, where $\eta(x)$ is an arbitrary smooth function that vanishes at the endpoints ($\eta(a) = \eta(b) = 0$) and ϵ is a small parameter.

The functional with the perturbed function is:

$$J(y+\epsilon\eta) = \int_a^b L(x, y(x)+\epsilon\eta(x), y'(x)+\epsilon\eta'(x))dx.$$

Expanding this to first order in ϵ, we have:

$$J(y+\epsilon\eta) \approx J(y) + \epsilon\int_a^b \left(\frac{\partial L}{\partial y}\eta(x) + \frac{\partial L}{\partial y'}\eta'(x)\right)dx.$$

The first variation $\delta J(y;\eta)$ is then:

$$\delta J(y;\eta) = \int_a^b \left(\frac{\partial L}{\partial y}\eta(x) + \frac{\partial L}{\partial y'}\eta'(x)\right)dx.$$

To simplify the term involving $\eta'(x)$, we integrate by parts:

$$\int_a^b \frac{\partial L}{\partial y'}\eta'(x)dx = \left[\frac{\partial L}{\partial y'}\eta(x)\right]_a^b - \int_a^b \frac{d}{dx}\left(\frac{\partial L}{\partial y'}\right)\eta(x)dx.$$

Since $\eta(a) = \eta(b) = 0$, the boundary term vanishes, leaving:

$$\int_a^b \frac{\partial L}{\partial y'}\eta'(x)dx = -\int_a^b \frac{d}{dx}\left(\frac{\partial L}{\partial y'}\right)\eta(x)dx.$$

Substitute this result back into the expression for the first variation:

$$\delta J(y;\eta) = \int_a^b \left(\frac{\partial L}{\partial y} - \frac{d}{dx}\left(\frac{\partial L}{\partial y'}\right)\right)\eta(x)dx.$$

For $\delta J(y;\eta) = 0$ for all arbitrary functions $\eta(x)$, the integrand must be zero:

$$\frac{\partial L}{\partial y} - \frac{d}{dx}\left(\frac{\partial L}{\partial y'}\right) = 0.$$

It provides the necessary condition for $y(x)$ to be an extremum of the functional $J(y)$.

11.4 Preliminary of the Flow model

In this section, we focus on how to derive the flow model. Let's start with MLE first.

11.4.1 MLE

Maximum Likelihood Estimation (MLE) is a method used in statistics to estimate the parameters of a statistical model. It involves finding the parameter values that maximize the likelihood function, representing the probability of observing the given data under the specified model. Here we first present the formula of the distribution of variable x and its MLE as below:

$$q_\theta(x) = \int q(z)q(x \mid z)dx.$$
$$MLE = \max_\theta \mathop{\mathbb{E}}_{x\sim p(x)} \log q_\theta(x).$$

VAE formally addresses the MLE issue using ELBO while GAN leverages discriminator and minimax game. However, the Flow model directly does the integration.

11.4.2 Settings of Flow model

Assume the variable z satisfies $\mu(z) \sim \mathcal{N}(0, \mathrm{I})$, to remove randomness from variable x, we can represent $q(x \mid z)$ as follows:

$$q(x \mid z) = \delta(x - g(z)).$$

where g is a generator. Accordingly, we can equivalently get the following formula:

$$q_\theta(x) = \int \delta(x - g(z))d\mu(z).$$

Thus, we have $q_\theta \sim g_\# \mu$ where $g_\# \mu$ represents a new distribution that given map $g, g_\# \mu$ can be formed from the original one μ. In other words, we have $\mu \sim g_\#^{-1} q_\theta$. Here we can derive the target density function:

$$q_\theta(x) = \mu\left(g^{-1}(x)\right) \cdot \left|\frac{Dz}{Dx}\right|.$$

where g is invertible and Jacobin determinant $\left|\frac{Dz}{Dx}\right|$ is easy to compute.

11.5 NICE

NICE is the first flow model paper that made the Jacobin matrix into a triangle matrix. Specifically, assume matrix h can be embedded into two pairs like this:

$$h = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 + m_\theta\left(x_1\right) \end{pmatrix}.$$

Thus, we can obtain:

$$\frac{Dh}{DX} = \begin{pmatrix} I & \\ \frac{Dm_\sigma}{DX_1} & I \end{pmatrix} \Rightarrow \left|\frac{Dh}{DX}\right| = 1.$$

We can also compute the convertible process:

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} h_1 \\ h_2 - m_0\left(X_1\right) \end{pmatrix} = \begin{pmatrix} h_1 \\ h_2 - w_0\left(h_1\right) \end{pmatrix} = g^{-1}(h).$$

11.5.1 How to make the setting useful

First,

$$h_l \circ h_{l_{-1}} \cdots \circ h_0(x) = z.$$

where z is noise. Second, shuffle h_1 and h_2. Third, scale transforms. Then, we can derive the following equation from the last layer:

$$z = S \odot h_e, \Rightarrow \frac{Dz}{Dh_e} = \operatorname{diag}(S).$$

where $\odot$ is the Hadamard product.

11.5.2 Real-valued non-volume preserving transformation

We know that:

$$\left|\frac{Dz}{Dh_e}\right| \neq 1.$$

So, we can derive the final form:

$$h = \begin{pmatrix} x_1 \\ S\left(x_1\right) \odot x_2 + m\left(x_1\right) \end{pmatrix} \Rightarrow \frac{Dh}{DX} = \begin{pmatrix} I & \\ \frac{DS}{DX_1} \odot X_2 + \frac{Dm}{DX_1} & \operatorname{diag}(S) \end{pmatrix}.$$

where $S\left(x_1\right) = \exp(\log S)$ and $\log S = \mathrm{NN}\left(x_1\right)$. A positive S can be generated by equation $S\left(x_1\right) = \exp(\log S)$. In this way, we can easily compute the Jacobin determinant.

11.5.3 How to use convolution as a trick

First, we can shuffle the channel but not X, Y. Second, we can use squeezing to increase channel size. For example,

$$h \times w \times c \rightarrow \frac{h}{2} \times \frac{w}{2} \times 4c.$$

11.5.4 How to sample after obtaining the models

Assume $z \sim \mathcal{N}(0, \mathbf{I})$, we can sample $z = h_l$, with $g_0^{-1} \circ g_1^{-1} \circ \cdots g_l^{-1}(z)$.

11.5.5 Multi-level Flow

Given matrix X, we can design multi-level flows:

$$X \xrightarrow{flow_1} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \xrightarrow{flow_2} \begin{pmatrix} z_1 \\ z_3 \\ z_4 \end{pmatrix} \xrightarrow{flow_3} \begin{pmatrix} z_1 \\ z_3 \\ z_5 \end{pmatrix}$$

Notably, we cannot assume $\begin{pmatrix} z_1 \\ z_3 \\ z_5 \end{pmatrix} \sim N(0, \mathbf{I})$. We have:

$$\begin{aligned} P\left(z_1, z_3, z_5\right) &= P\left(z_1 \mid z_3, z_5\right) \cdot P\left(z_3 \mid z_5\right) \cdot P\left(z_5\right) \\ &= P\left(z_1 \mid z_2\right) \cdot P\left(z_3 \mid z_4\right) \cdot P\left(z_5\right). \end{aligned}$$

where $\sigma\left(z_5\right) = \sigma\left(z_4\right), \sigma\left(z_3, z_5\right) = \sigma\left(z_2\right), z_1 \sim \mathcal{N}\left(\mu\left(z_2\right), \Sigma\left(z_2\right)\right), z_3 \sim \mathcal{N}\left(\mu\left(z_4\right), \Sigma\left(z_4\right)\right)$ and $z_5 \sim \mathcal{N}(\mu, \Sigma)$. And μ and Σ are tunable. Thus:

$$\begin{aligned} P\left(z_1, z_3, z_5\right) &= P\left(z_1 \mid z_2\right) \cdot P\left(z_3 \mid z_4\right) \cdot P\left(z_5\right) \\ &= \mathcal{N}\left(\mu\left(z_2\right), \Sigma\left(z_2\right)\right) \cdot \mathcal{N}\left(\mu\left(z_4\right), \Sigma\left(z_4\right)\right) \cdot \mathcal{N}(\mu, \Sigma). \end{aligned}$$

Here is the generation process:

$$\begin{aligned}
&z_5 \sim \mathcal{N}(\mu, \Sigma), z_4 = \text{ flow }_3^{-1}\left(z_5\right), \\
&z_3 \sim \mathcal{N}\left(\mu\left(z_4\right), \Sigma\left(z_4\right)\right), z_2 = \text{ flow }_2^{-1}\begin{pmatrix} z_3 \\ z_4 \end{pmatrix}, \\
&z_1 \sim \mathcal{N}\left(\mu\left(z_1\right), \Sigma\left(z_2\right)\right), \\
&X = \text{ flow }_1^{-1}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}.
\end{aligned}$$