Information Theory

Additional mini-project in deeplearning math

Kullback-Leibler Divergence (KL 발산)

소프트웨어 꼰대 강의

노기섭 교수

(kafa46@cju.ac.kr)

Course Overview

Topic	Contents
01. Orientation 오리엔테이션	Motivations & Course introduction 동기부여, 과정 소개
02. Information 정보	What is the information? Concept & definition 정보란 무엇인가? 개념과 정의
03. Information Entropy 정보 엔트로피	Concepts, notation, and operations on information entropy 정보 엔트로피의 개념, 표기, 연산
04. Entropy in Deeplearning 딥러닝에서의 엔트로피	How to apply the information entropy into Deeplearning? 어떻게 정보 엔트로피를 딥러닝에 적용하는가?
05. Entropy Loss 엔트로피 손실	Loss function using entropy, BCE, and cross entropy 엔트로피를 이용한 손실 함수, BCE, 크로스 엔트로피
06. KL Divergence KL 발산	Concept & definition of KL divergence KL 발산의 개념과 정의
07. Summary & Closing 요약 및 마무리	Summary & closing on this project, 'Information Theory' 정보 이론 요약 및 마무리

Introduction to the inventors

In mathematical statistics, the Kullback–Leibler (KL) divergence (also called relative entropy and I-divergence[1]), denoted $D_{KL}(P||Q)$, is a type of statistical distance.

A measure of how one probability distribution P is different from a second, reference probability distribution Q.

Kullback, S.; Leibler, R.A. (1951).
"On information and sufficiency".
Annals of Mathematical Statistics. 22 (1): 79–86.
doi:10.1214/aoms/1177729694

ON INFORMATION AND SUFFICIENCY

BY S. KULLBACK AND R. A. LEIBLER

The George Washington University and Washington, D. C.

1. Introduction. This note generalizes to the abstract case Shannon's definition of information [15], [16]. Wiener's information (p. 75 of [18]) is essentially the same as Shannon's although their motivation was different (cf. footnote 1, p. 95 of [16]) and Shannon apparently has investigated the concept more completely. R. A. Fisher's definition of information (intrinsic accuracy) is well known (p. 709 of [6]). However, his concept is quite different from that of Shannon and Wiener, and hence ours, although the two are not unrelated as is shown in paragraph 2.

Solomon Kullback

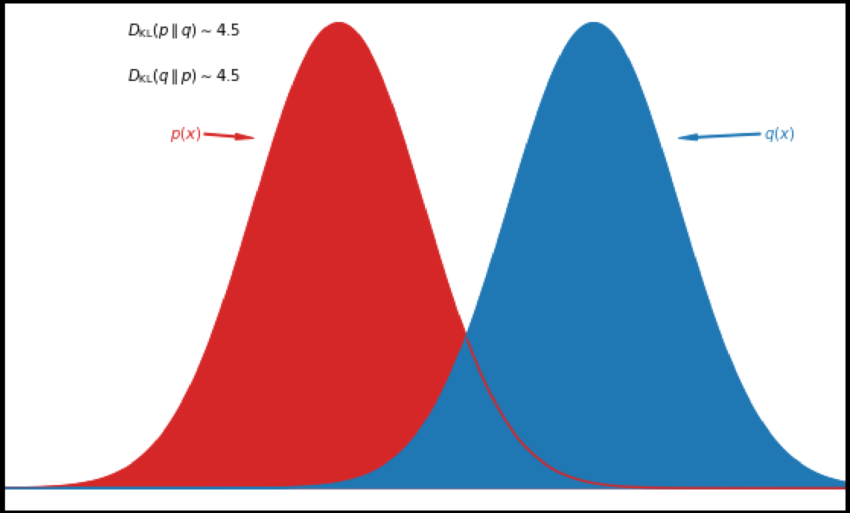
Born	April 3, 1907 Brooklyn, New York
Died	August 5, 1994 (aged 87) Boynton Beach, Florida
Citizenship	American
Alma mater	City College of New York (B.A., 1927; M.A., 1929) George Washington University (Ph.D., Mathematics, 1934)
Known for	Work in Information theory, Kullback–Leibler divergence
Scientific career	
Fields	cryptanalysis, mathematics, information theory
Institutions	George Washington University, National Security Agency

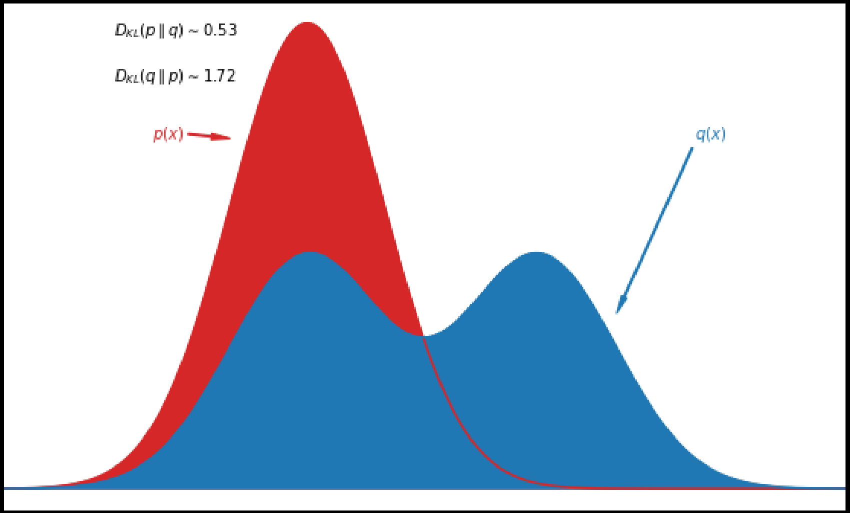
Richard Leibler

Born	March 18, 1914 Chicago, Illinois
Died	October 25, 2003 (aged 89) Reston, Virginia
Alma mater	Northwestern University (A.M., Mathematics) University of Illinois (Ph.D., Mathematics, 1939)
Known for	Kullback–Leibler divergence
Scientific career	
Fields	cryptanalysis, mathematics
Institutions	United States Navy, Princeton University, National Security Agency, Institute for Defense Analysis

딥러닝에 자주 등장하는 KL Divergence

두 분포가 있을 경우, 그 차이를 어떻게 측정할까???





이미지 출처: https://datumorphism.leima.is/wiki/machine-learning/basics/kl-divergence/

두 분포의 차이를 측정할 수 있다면?

$Y(P_{Label})$의 확률 분포와 의 $\hat{Y}(P_\theta)$확률분포의

차이를 최소화 하도록 최적화 가능할 것

KL Divergence 정의

기준이 되는 분포

Definition

$$D(p||q) = \mathbb{E}_{p}\left[\log \frac{\boxed{p(X)}}{q(X)}\right] = \sum_{x \in \mathcal{X}} \boxed{p(x)} \times \log \frac{\boxed{p(x)}}{q(x)}$$

$, where \;\; p \;\&\; q \;\; are \;\; probability\ distribution$

$X \;\; is \;\; random \;\; variable$ and $0\log\frac{0}{0} = 0, \quad 0\log\frac{0}{q} = 0, \quad 0\log\frac{p}{0} = \infty$

분자와 분모를 바꿔서 표현해도 무방합니다 ^^.

$$D(p||q) = \mathbb{E}_{p}\left[\log \frac{p(X)}{q(X)}\right] = -\sum_{x \in \mathcal{X}} p(x) \times \log \frac{q(x)}{p(x)}$$

Distance vs. Divergence

Note:
In general

$$D(p||q) \neq D(q||p)$$

거리 개념을 사용하는 것에
약간의 이견이 존재

Euclidean distance 관점에서는 틀린 말

일반적 거리 관점에서는 맞는 말

유명한 책 'Element of Information Theory' 그리고
Online Wiki 에서는 KL Divergence를
KL distance 라고 표현

In information geometry, a divergence is a kind of statistical distance: a binary function which establishes the separation from one probability distribution to another on a statistical manifold.

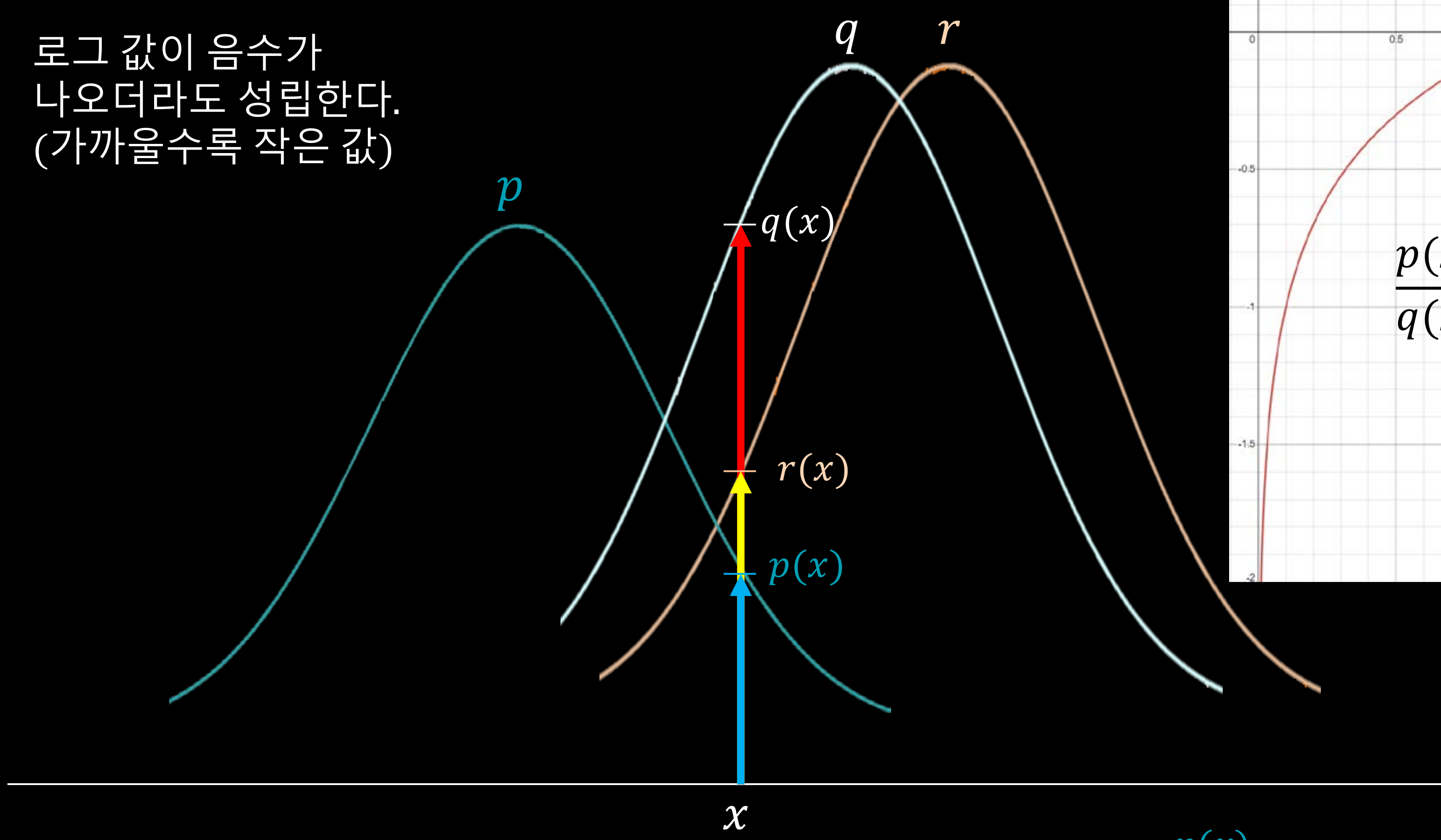
The simplest divergence is squared Euclidean distance (SED), and divergences can be viewed as generalizations of SED.

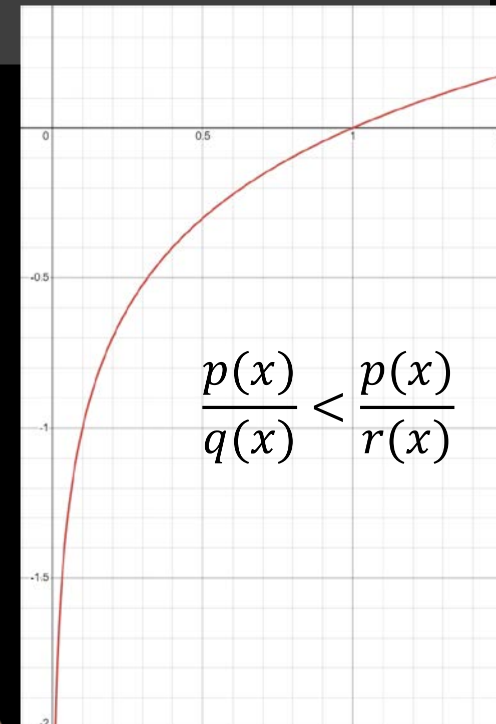
The other most important divergence is relative entropy (also called Kullback–Leibler divergence)

Source: https://en.wikipedia.org/wiki/Divergence_(statistics)

손으로 구해보는 KLD (1/2)

로그 값이 음수가
나오더라도 성립한다.
(가까울수록 작은 값)





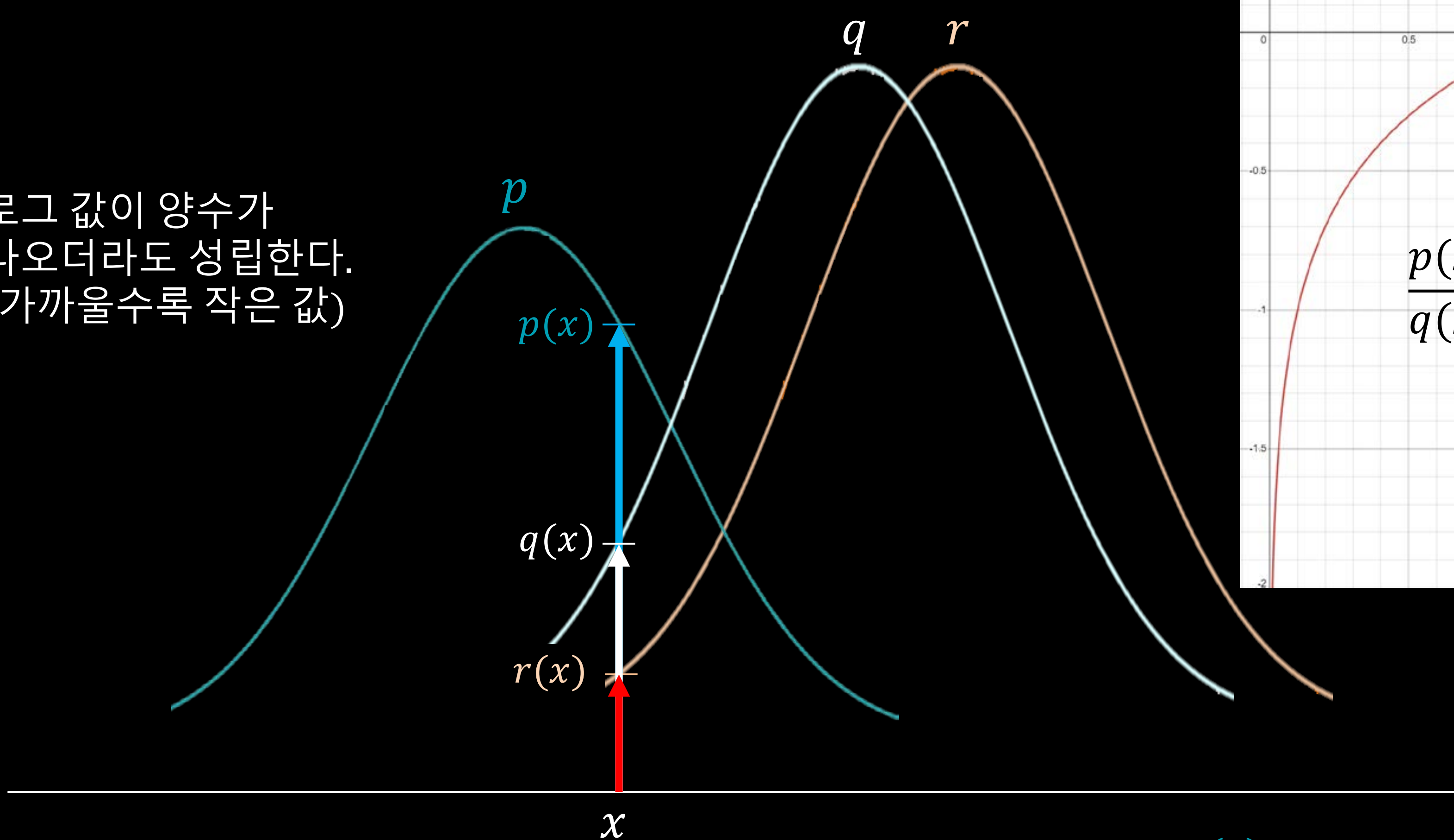
$$\frac{p(x)}{q(x)} < \frac{p(x)}{r(x)}$$

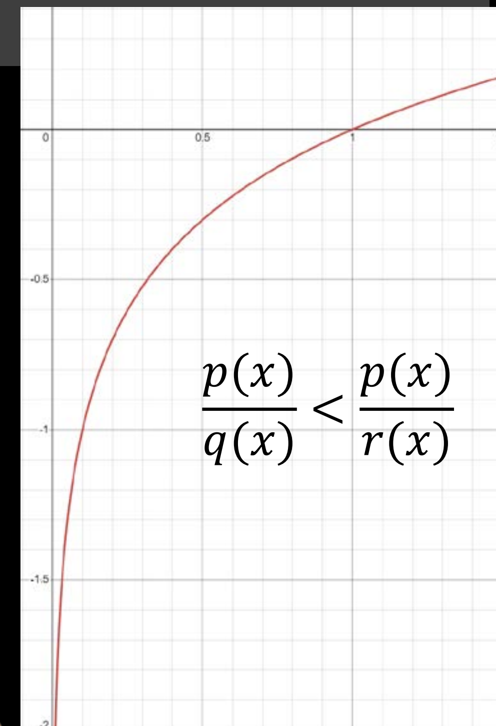
$$D(p||q) = p(x) \times \log\frac{p(x)}{q(x)}$$

$$D(p||r) = p(x) \times \log\frac{p(x)}{r(x)}$$

손으로 구해보는 KLD (2/2)

로그 값이 양수가
나오더라도 성립한다.
(가까울수록 작은 값)





$$\frac{p(x)}{q(x)} < \frac{p(x)}{r(x)}$$

$$D(p||q) = p(x) \times \log \frac{p(x)}{q(x)}$$

$$D(p||r) = p(x) \times \log \frac{p(x)}{r(x)}$$

KL Divergence의 범위는?

로그 합 부등식 (log sum inequality)

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

증명 (proof)

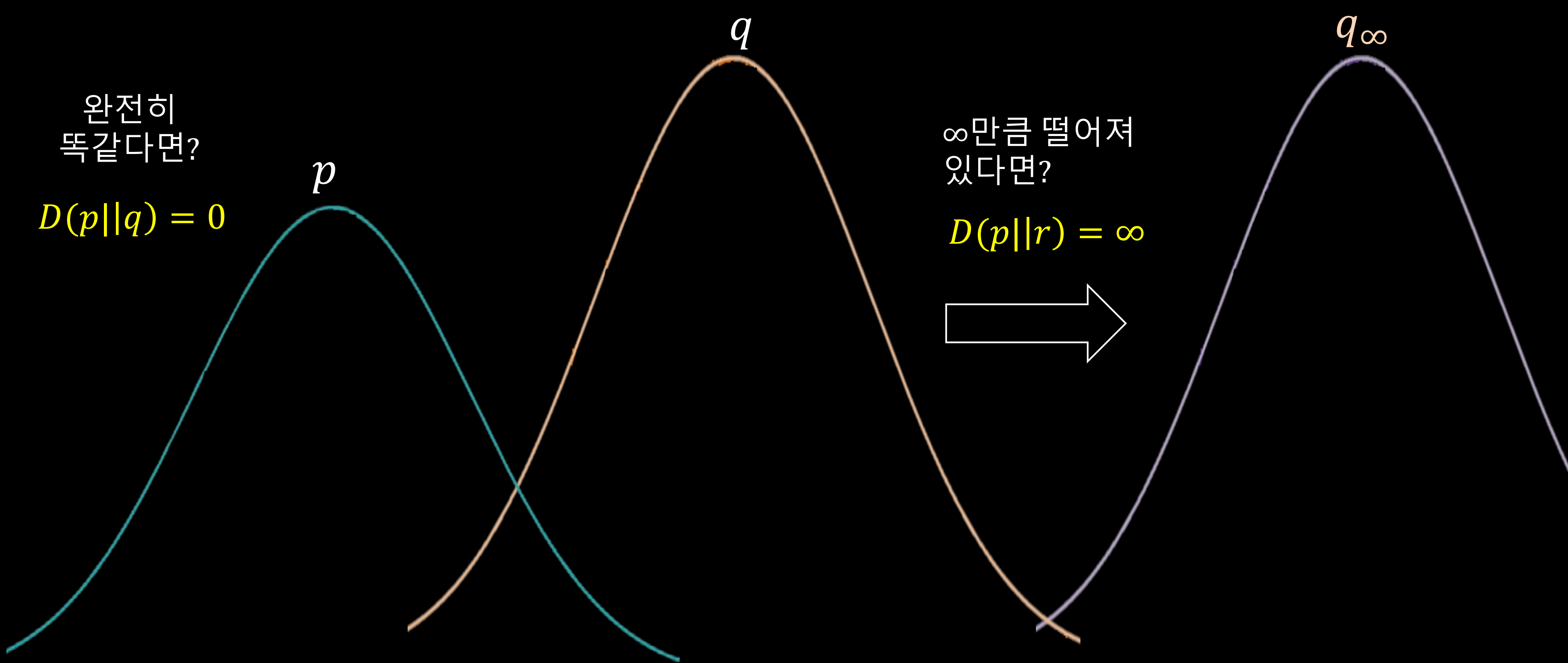
$$D(p||q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)}$$

$$\geq \left(\sum_{i=1}^{n} p_i\right) \log \frac{\sum_x p(x)}{\sum_x q(x)} = 1 \times \log \frac{1}{1} = 0$$

$Therefore,$

$$D(p||q) \geq 0$$

KL Divergence 개념에 대한 이해



$$D(p||r) = p(x) \times \log \frac{p(x)}{q(x)}$$

KLD in Deeplearning Optimization

$$L(\theta) = D(Y||\hat{Y}) = \mathbb{E}_Y\left[\log \frac{Y}{\hat{Y}}\right]$$

$\hat{Y}$ 은 softmax를 통과한 확률분포라고 가정

데이터셋 수집 ⇨ $Dataset = \{(x_i, y_i)\}_{i=1}^{n}$

Monte Carlo 근사

이전강의를 참고해 주세요 ^^
"[Probability]_06. 샘플링 표현에 대한 이해와 몬테 카를로 근사"
https://youtu.be/nw_tVBCw0Z8

$$L(\theta) \approx \frac{1}{n} \times p \sum_{i=1}^{n} \log \frac{y_i}{\hat{y}_i}$$

장점?

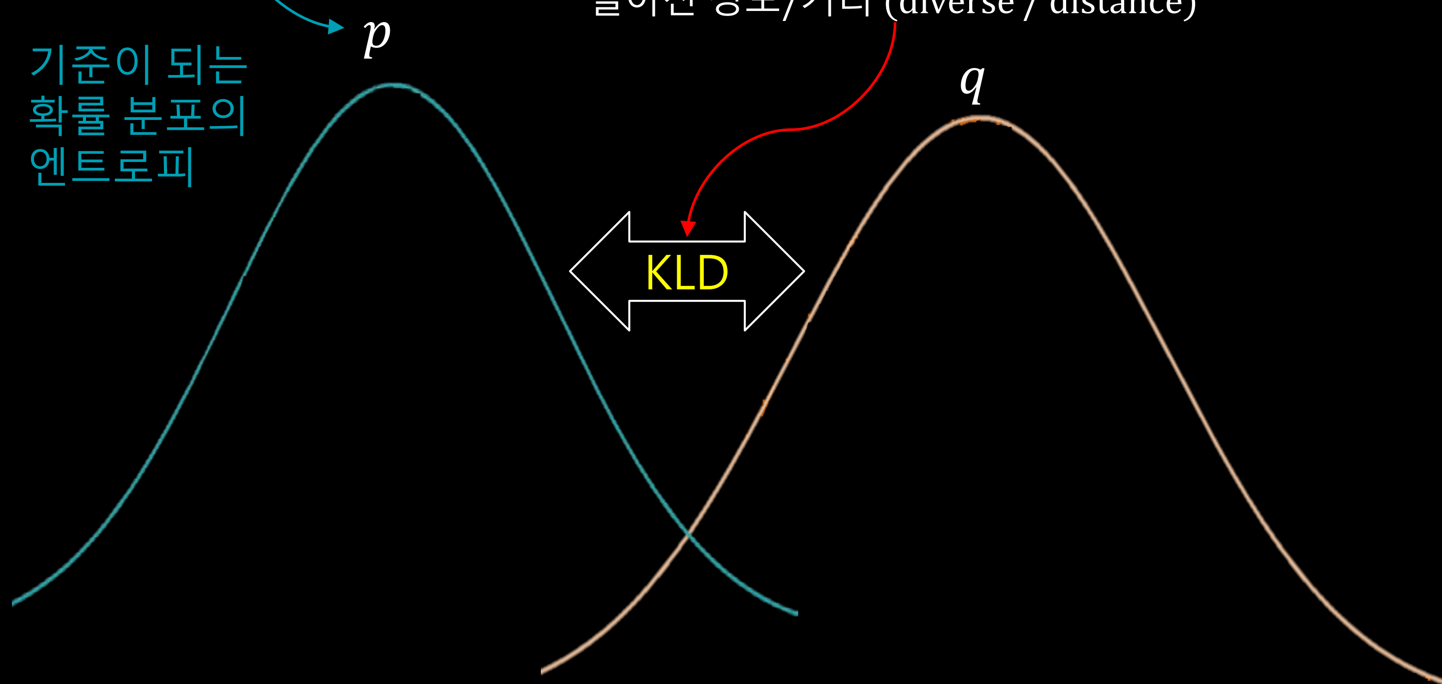
예측 오차를 줄이면서 전체적인 확률 분포도 같아지도록 학습할 수 있습니다.

KLD와 Cross Entropy 관계?

$$H(p,q) = \boxed{H(p)} + \boxed{D(p||q)}$$

비교 대상이 되는 확률 분포와
떨어진 정도/거리 (diverse / distance)

기준이 되는
확률 분포의
엔트로피



수고하셨습니다 ..^ ^..