Variational Inference

swear013@gmail.com

norman3.github.io

Information

- 아이디어는 간단하다.
- 복잡한 분포(distribution)를 좀 더 간단한 형태의 분포로 근사하자는 것.
- 물론 근사 분포를 사용하는 모델은 이것 말고도 많다.
  - 이런 스타일을 사용하는 모델 중 하나라고 생각하면 된다.
- Variational 까지 바로 달려가기 위해 이론들을 최대한 압축하여 설명함.

Basic Theory

Probability

$$p(x)$$

- “주사위의 한 면이 나올 확률은 1/6 이야”
  - 전지적 관점에서의 확률
- “주사위 굴리기를 10만번 반복한 결과 한 면이 나올 확률은 각각 1/6 이고 오차 비율은 XX야.”
  - 빈도적 관점에서의 확률
- “내일 지구가 멸망할 확률은 0.00000000000000001 이야”
  - 한번도 일어나지 않은 사건에 대한 예측 확률
- “전문가의 예측에 따르면 내년에 한화가 우승할 확률은 내일 지구가 멸망할 확률과 같아.”
  - 주관에 따른 확률

Probability (Cont'd)

$$p(x)$$

- 기계 학습에 확률 이론을 도입하기 위해서는?
  - 애매모호한 확률 개념 대신 아주 명확한 정의를 통해 확률을 사용할 수 있어야 한다.

- **확률은 함수다. <= 이 사실만 기억하면 된다.**
  - 실수 벡터를 입력으로 받아 실수 값을 출력.
  - 아무 함수는 아니고 특정한 **"제약"** 을 가진 함수

- 확률 관련 좋은 자료 (단점은 영문임)
  - http://cs229.stanford.edu/section/cs229-prob.pdf

Probability (Cont'd)

- 확률 함수 (Probability function ??)
  - 실수 벡터를 입력으로 받아 실수 값을 출력.
  - 입력 실수 벡터는 랜덤 변수.
  - 합이 1이거나 (PMF) 적분이 1 (PDF).
  - 출력 실수 값은 0보다 커야 한다.

PMF

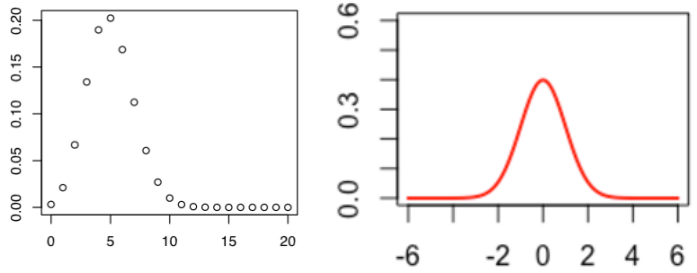
$p(x_i) \geq 0$ $\sum_i^N p(x_i)=1$

R.V

PDF

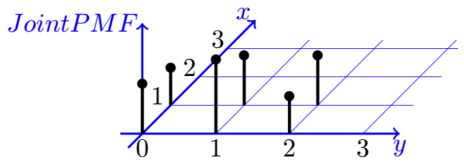
$p(x_i) \geq 0$ $\int_{-\infty}^{\infty} p(x_i) = 1$

- 해야 할 이야기는 많지만 시간이 없는 관계로 생략한다.
  - PMF와 PDF의 차이를 알고 있어야 한다.



주요 확률식 (2개의 랜덤 변수 사이)

$p_{x,y}(x,y)$



$p_{x,y}(x,y) = p_x(x)p_y(y)$

두 R.V. 가 독립인 경우

$p_{x,y}(x \mid y)$

어쨌거나 x와 y 에 관한 함수이긴 한데
y 가 결정되어야 뭘 알아낼 수 있다.

$P \quad (x = 3|y) \quad P \quad (x = 3|y = 0)$

$P_x(x|y = 1) \quad P \quad (x = 3|y = 1)$

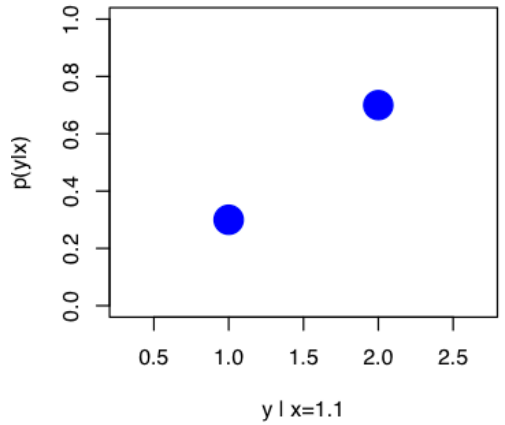
$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

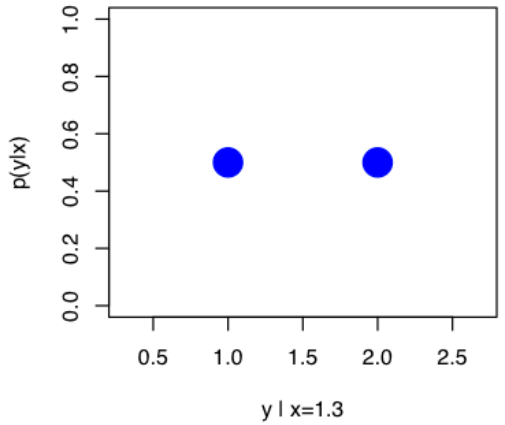
조건부 분포

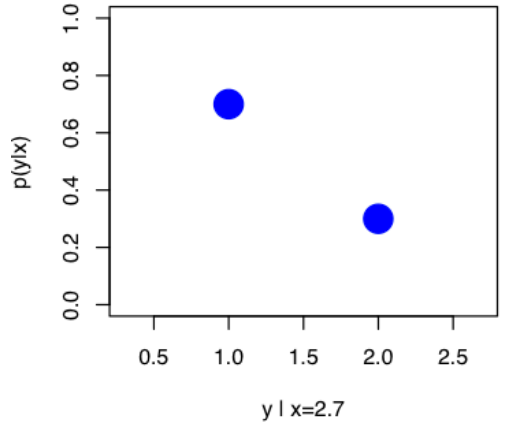
$$p(y|x)$$

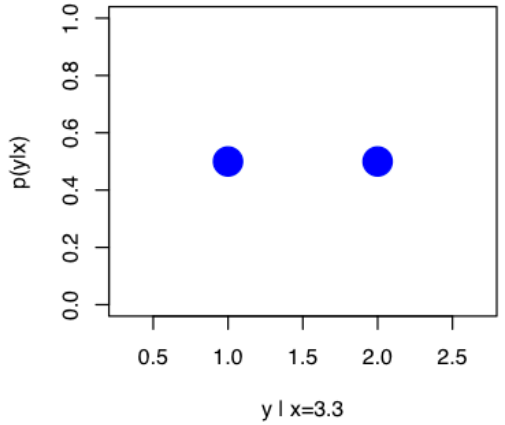
$$p(y|x = 1.1)$$

$$p(y = 1|x = 1.1)$$





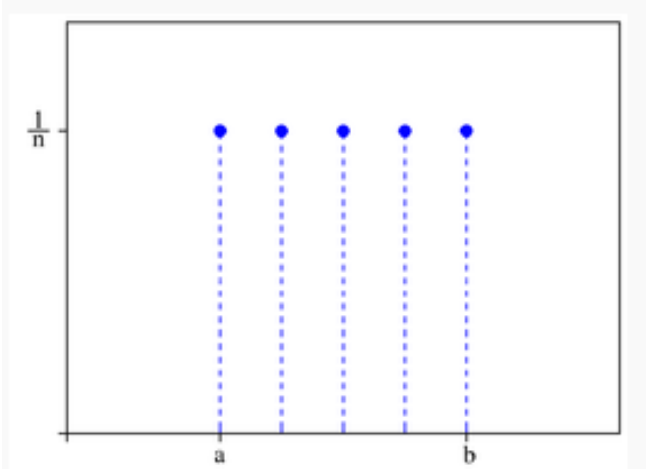




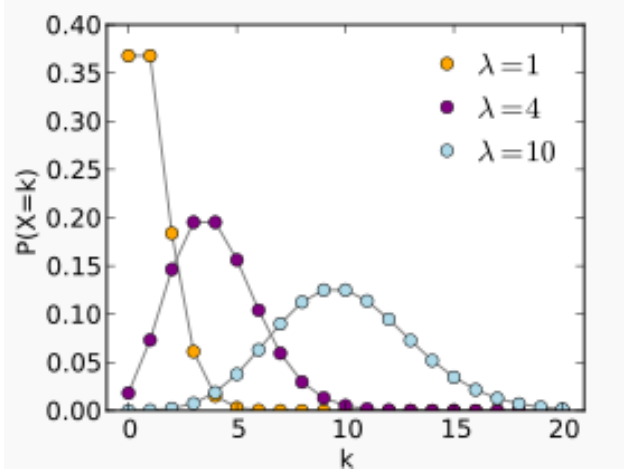
Probability Distribution with Parameter

- 분포(distribution)는 사실 그냥 확률 함수와 같은 개념이라 생각하면 된다.
  - 어떤 경우에는 파라미터를 가진 확률 함수를 나타내는데 사용되기도 한다.
  - 파라미터(parameter)를 가진 확률 함수들은 보통 파라미터를 결정 하기만 하면, 해당 확률 함수의 모양을 결정 지을 수 있다. (인터넷을 찾아보자)
    - Ex) 정규분포, 포아상분포, 이항분포, 베타분포, 감마분포, 디리슈레분포 등등

$$p(x;\theta)$$

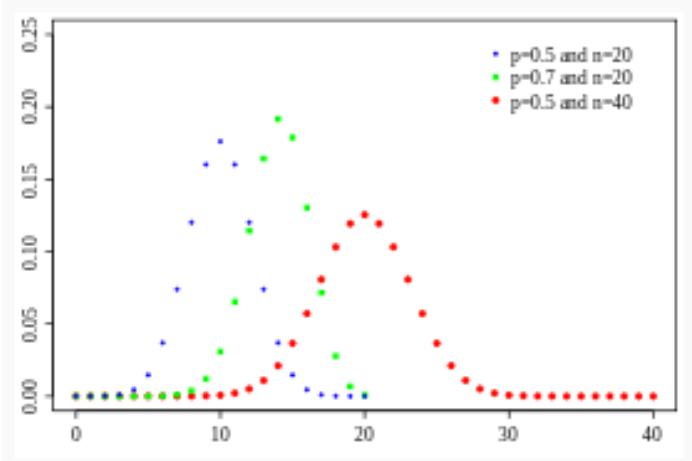


이산 균등분포 $\frac{1}{n}$



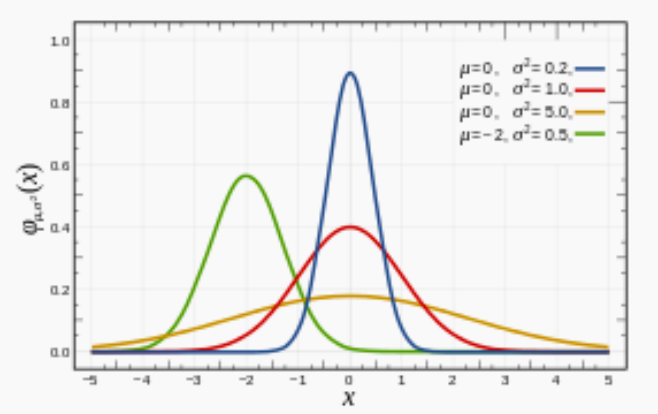
포아상분포

$$f(n;\lambda) = \frac{\lambda^n e^{-\lambda}}{n!},$$

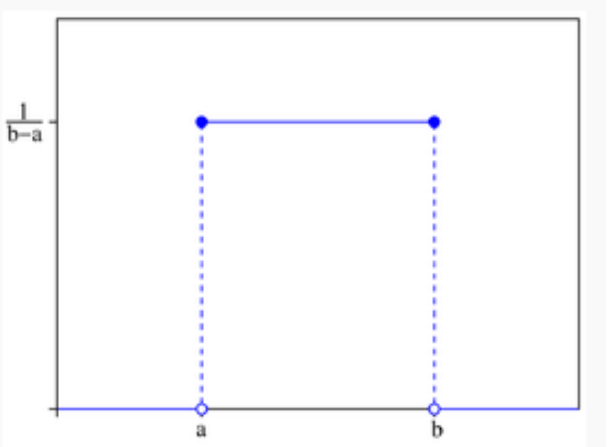


이항분포

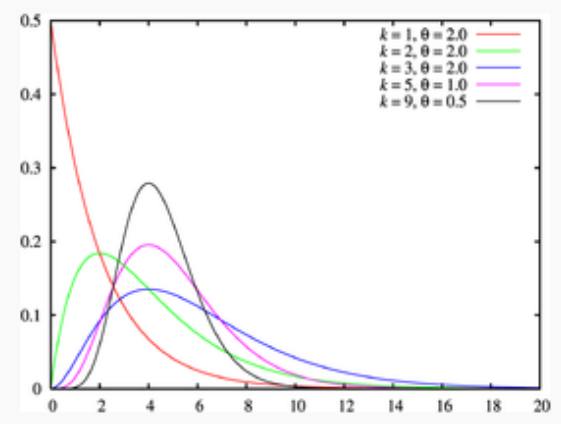
$$\Pr(K = k) = f(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$



$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



연속 균등분포

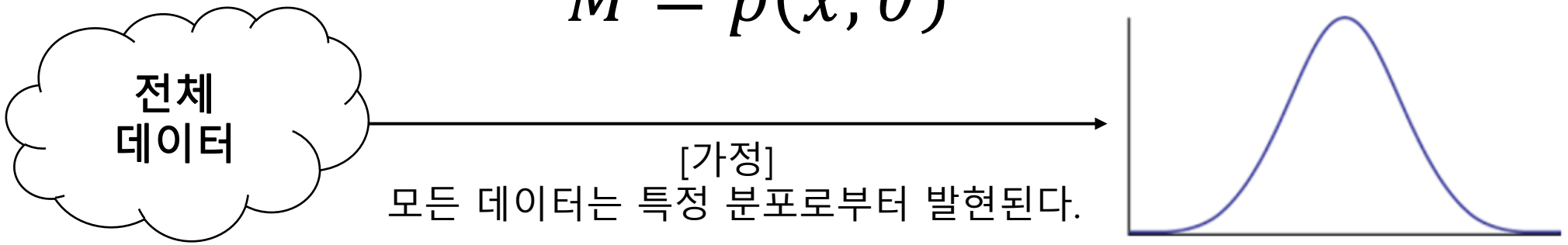


감마분포

$$f(x;k,\theta) = x^{k-1}\frac{e^{-x/\theta}}{\theta^k\,\Gamma(k)} \text{ for } x > 0$$

MLE (Maximum Likelihood Estimation)

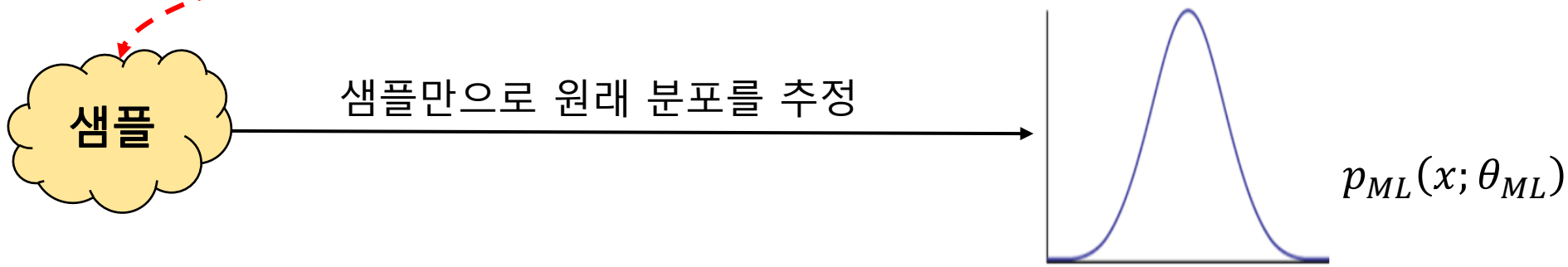
$$M = p(x;\theta)$$



샘플 또한 p(x)로부터 발현된 데이터이므로,
이 성질을 이용하여 적절한 L 함수를 설계한다.

$L(\theta;x) = argmax_\theta p(D;\theta)$

이런 이유로 likelihood 라는 이름이 생겼다.

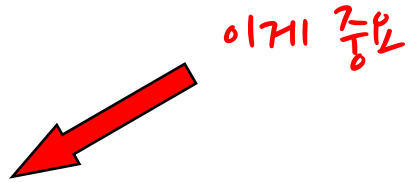


확률을 바라보는 두 관점

- Frequentist
  - 확률 : 빈도로 주어지는 확률
  - 파라미터는 알지 못하지만 고정된 상수
    - (unknown but fixed)
  - 많은 시행을 통해 좋은 추정을 얻을 수 있다.

$$L(\theta) = argmax_{\theta} p(D; \theta)$$

- Bayesian
  - 확률 : 믿음의 정도
  - 파라미터도 랜덤 변수가 될 수 있다.
  - 주어진 데이터로 좋은 추정을 얻어야 한다.

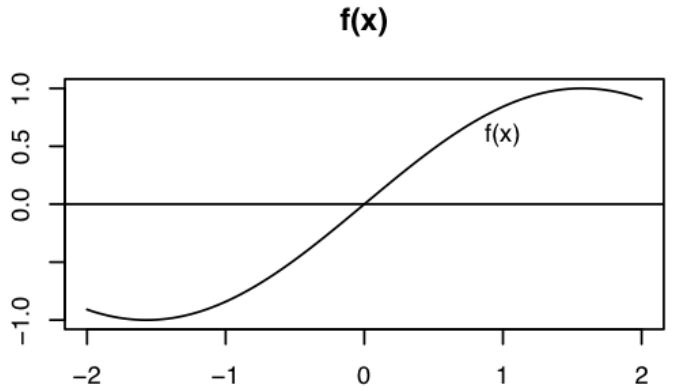


$$L(\theta) = argmax_{\theta} p(\theta|D)$$

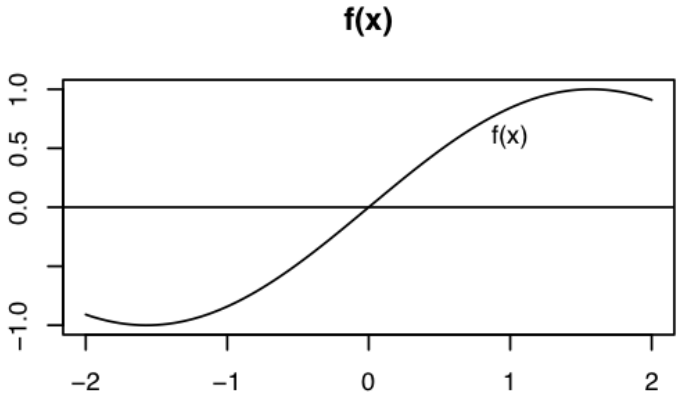
Expectation

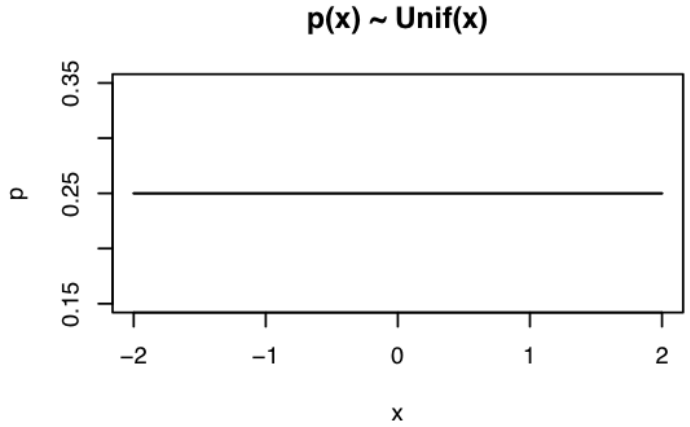
$$E_x\ [x] = \int xp(x)dx\ \ ,\ E_x\ [f] = \int p(x)f(x)dx$$

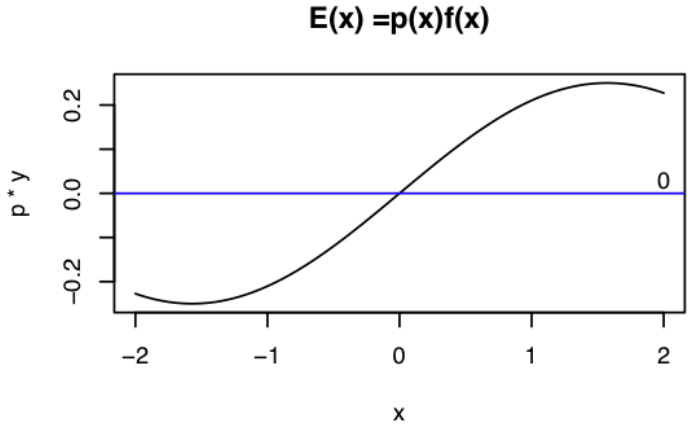
- 보통 값에 대한 기대값(평균)만을 생각하는 경우가 많다. ($E_x\ [x]$)
- 하지만 정말 중요한 것은 함수에 대한 기대값.
  - 원래 표준어로는 기댓값이 맞지만 어색하므로 기대값이라 하자.
    - 아래 함수의 기대값은 얼마인가?

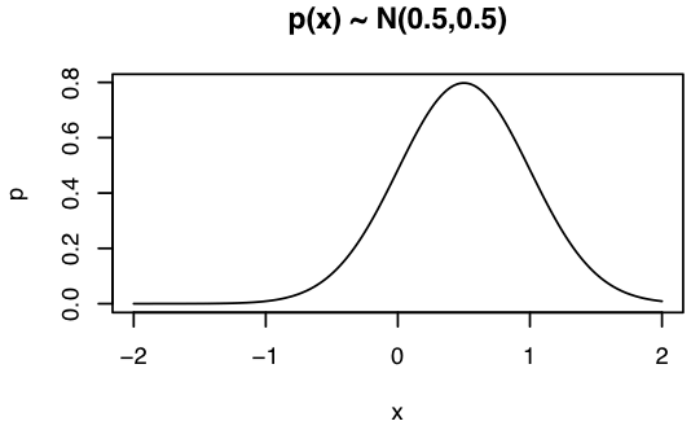


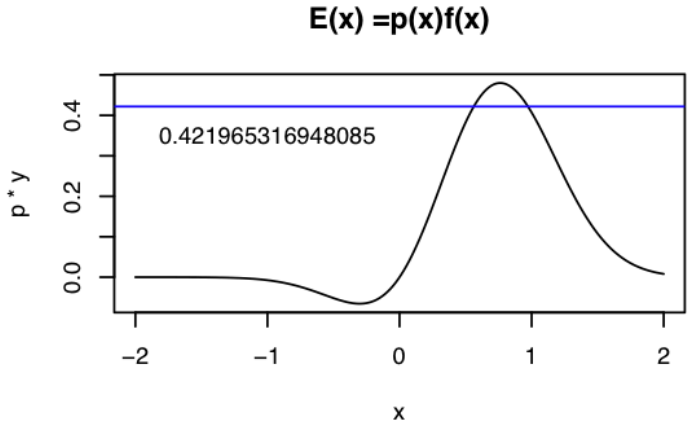
Expectation (Cont'd)



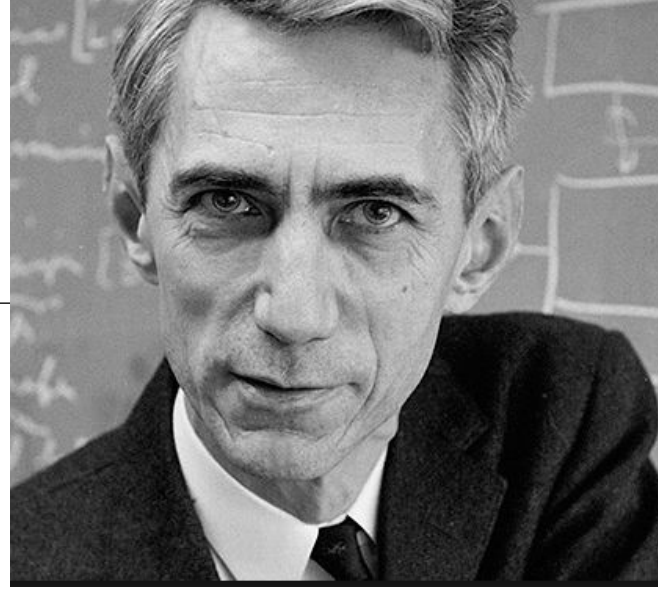








Information



- Shannon 횽아가 정의한 개념이다.
- 간단하게 생각하자면 그냥 “놀람”의 정도를 수치화.
  - 너무 뜬금없다 생각말고 사람이 만든 추상적 척도 개념이라 생각하자. 주가 지수 같은 그런 거..
- 정말 중요한 점은 확률 함수를 이용하여 정의된다는 것.
- 사실은 이를 증명하는 식은 무척이나 신비롭고 놀라운 이야기.
  - 얼마나 위대한 발견인지 볼츠만 횽아는 이 식을 자기 묘비에 썼다.
  - 하지만 우리는 이걸 다 볼 필요는 없고 **느낌적 느낌**만 알고 대충 넘어가도 된다.

$$h(x) = -\log\{p(x)\}$$

Information (Cont'd)

- 로또 1등에 대한 정보량 (숫자 6개)

$$h(x) = -\log_2\left(\frac{1}{8{,}145{,}060}\right) \cong 23$$

- 로또 5등에 대한 정보량 (숫자 3개)

$$h(x) = -\log_2\left(\frac{1}{45}\right) \cong 2$$

- 주사위 눈금이 1이 나올 확률에 대한 정보량

$$h(x) = -\log_2\left(\frac{1}{6}\right) \cong 2.6$$

Entropy

- Entropy란?
  - 하나의 계(system)가 가지는 **평균** 정보량.

$$H[x] = -\sum_{x} p(x)\log_2 p(x)$$

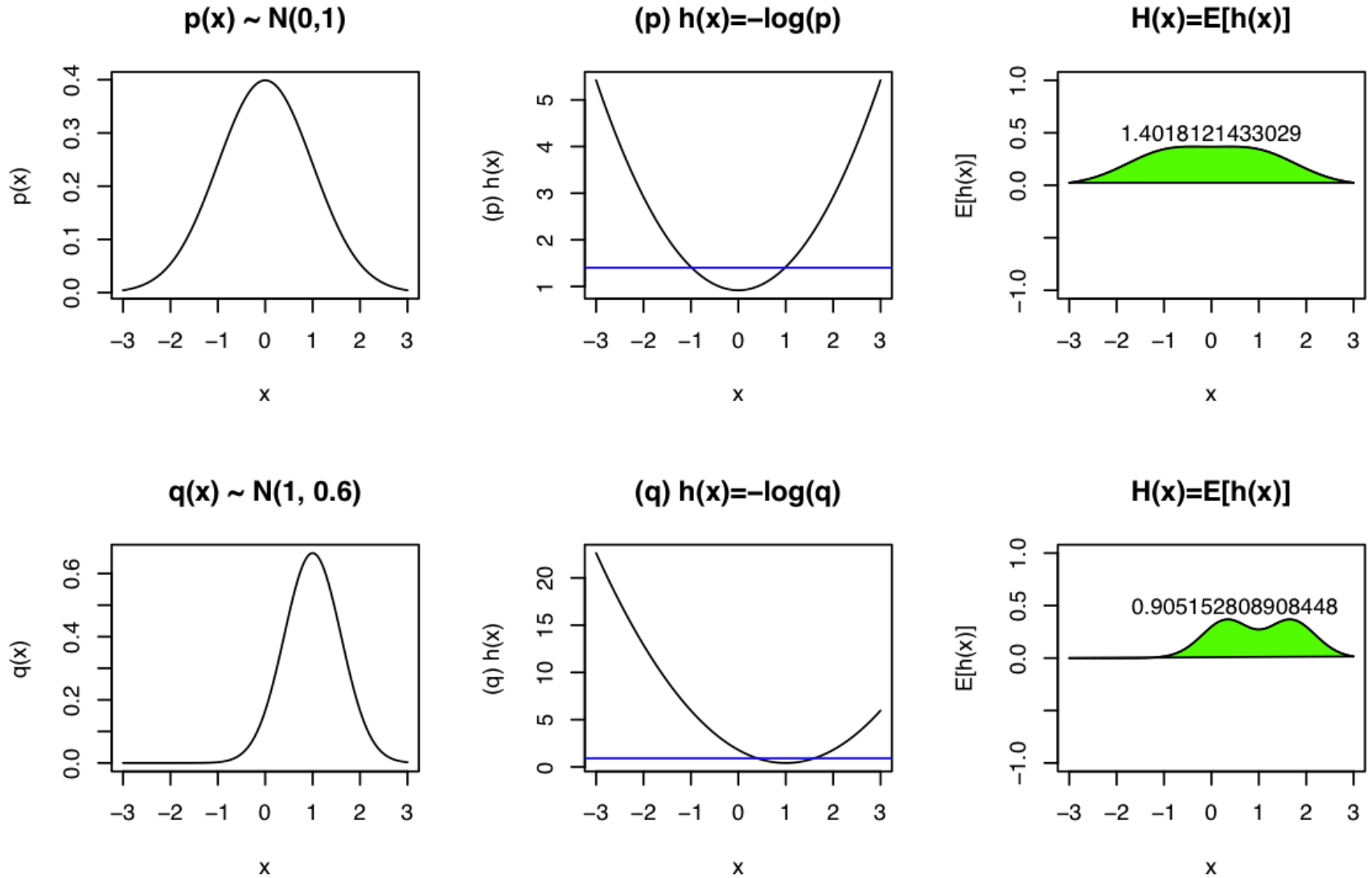
  - 이렇게 이야기하면 좀 멋져 보임. 생각해보면 별 의미 없는 말인데…
  - 어쨌거나 이런 값을 이용하면 “계 vs 계” 정보량 싸움을 붙일 수 있다.
  - 참고로 함수에 대한 평균 정의는 다음과 같다.

$$\mathrm{E}[f] = \sum p(x)f(x)$$

$$\mathrm{E}[f] = \int p(x)f(x)dx$$

  - Entropy 가 참 놀라운게 지수가 2인 로그를 사용하는 경우,
    - 우리가 사용하는 컴퓨터의 저장 단위인 bit 와 물리적 단위를 일치시킬 수 있다.
    - 쓸모가 아주 많아지게 된다. 평균 정보량 계산 등.

Entropy (Cont'd)



KL-divergence

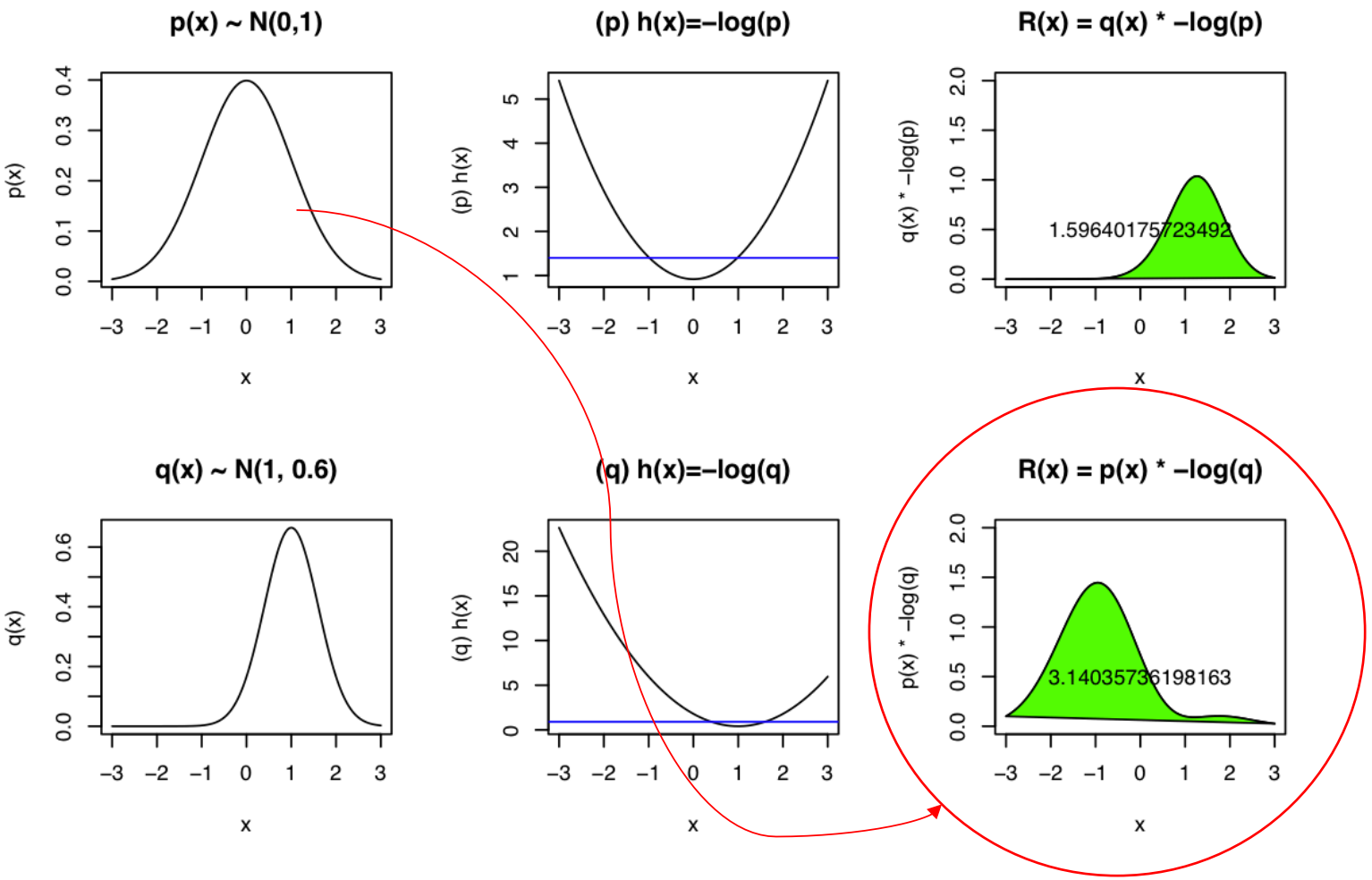
- 사람들이 대부분 제대로 모르면서 아는 척하는 개념. (발표자도 잘 모름)

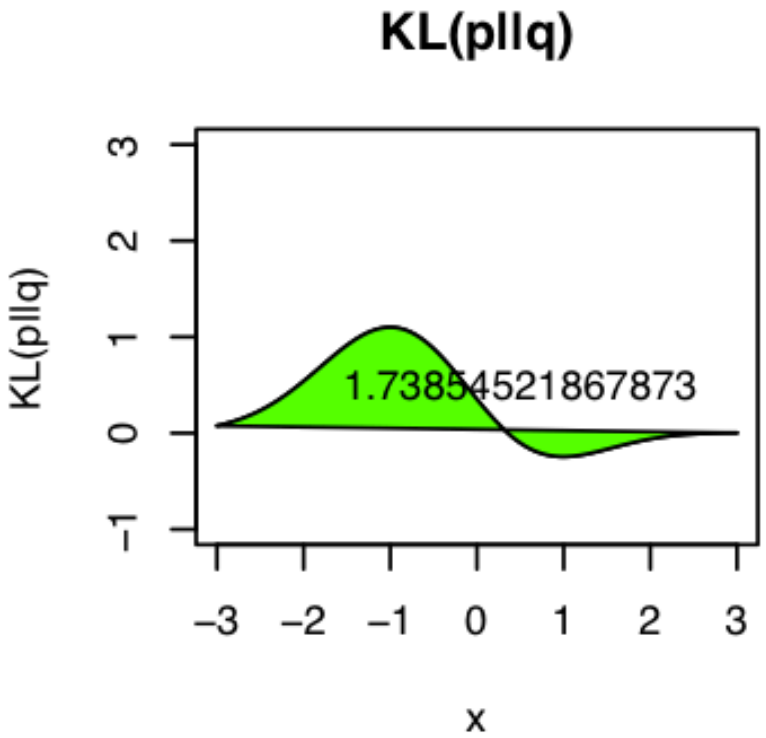
$$KL(p||q) = -\int p(\mathbf{x}) \ln q(\mathbf{x}) d\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x}\right) = -\int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} d\mathbf{x}$$

- 정의
  - P 라는 확률 분포로 부터 발생한 데이터를,
  - Q 라는 확률 분포에서 나왔다고 가정했을 경우
  - 이로 인해 발생되는 추가 정보량을 KL-divergence 이라고 한다… 역시 어렵다.
  - (잊지말자)
    - P와 Q가 같으면 KL 값은 0. 서로 다르면 KL값은 0보다 크다. 클수록 차이가 크다.

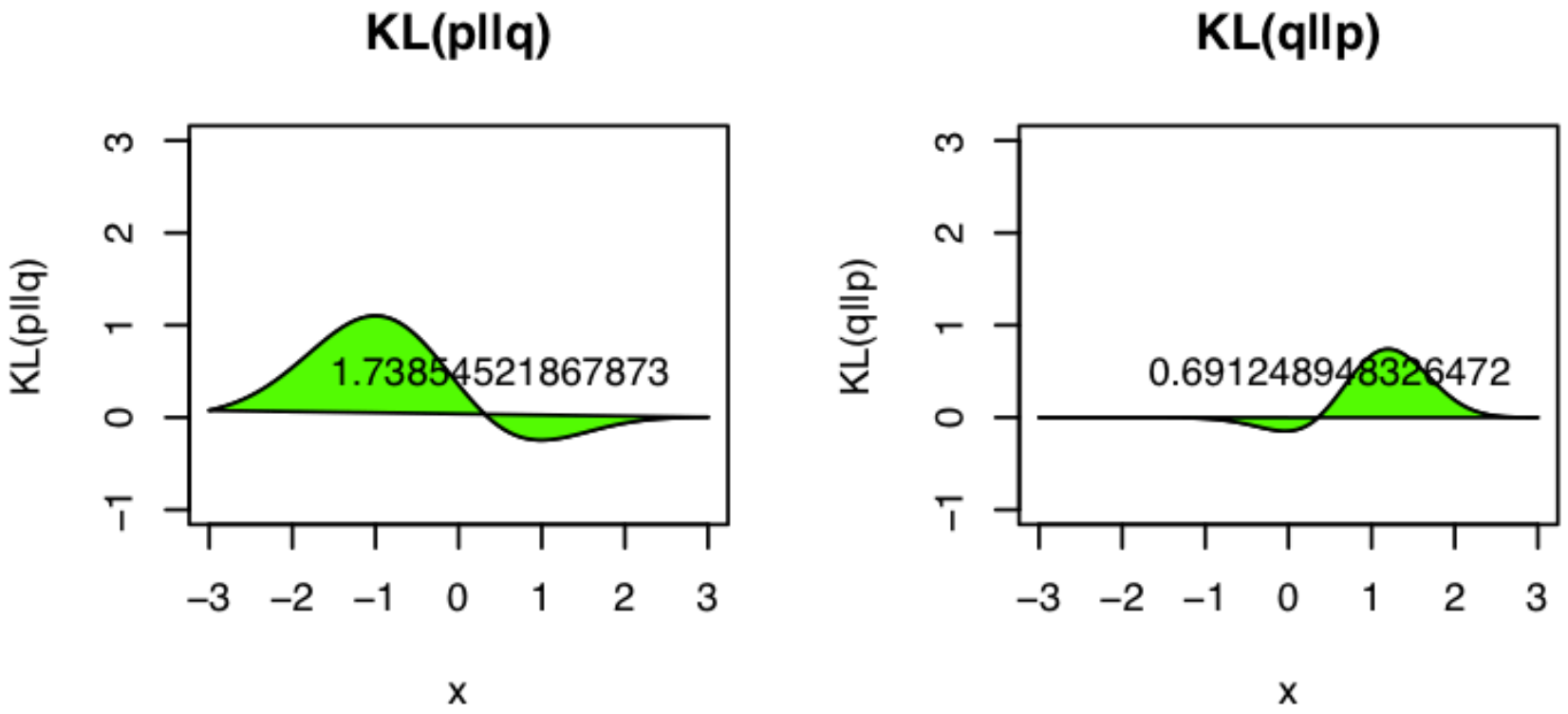
KL-divergence (Cont'd)

$$KL(p||q) = -\int p(\mathbf{x}) \ln q(\mathbf{x}) d\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x}\right)$$





KL-divergence (Cont'd)



Asymmetric

KL-divergence (Cont'd)

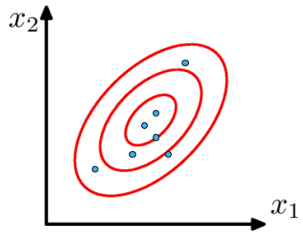
- 주 용도?
  - 우리가 원래의 분포 P 를 모르는 상태에서 샘플은 P 로부터 얻어진 상황이라면,
  - 어떤 Q라는 분포 함수를 도입하여 마치 이걸 P 인양 막 쓴다고 하자.
  - 그럼 이런 상황으로 인해 발생되는 오차율을 KL 값을 이용하여 상대 비교가 가능함.
  - 예를 들어 Q1, Q2 를 가정하고 각각 KL을 P에 대해 구해보니 Q1이 더 작다.
  - 그러면 Q1 이 Q2 보다 P에 더 가까운 모양이라고 고려할 수 있다.

EM (Expectation-Maximization)

갑작스럽지만 바로 EM 알고리즘을 살펴도록 하자.

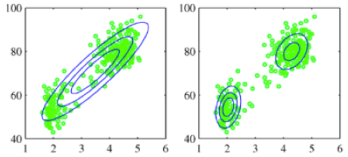
Mixture Distribution

- 데이터가 아주 귀하던 시절에는 적은 수의 데이터를 표현하는 적절한 수단이 필요했다.
  - 이를 위해 기초적인 분포(distribution)들을 널리 사용했다.
    - 정규분포, 균등분포, 포아상 분포, 스튜던트-t 분포, 감마 분포, 베타 분포 등등



- 데이터가 풍성해지자 단순한 분포로는 표현하기 어려운 데이터들이 생겨났다.

- 기존의 분포들을 서로 묶어 새로운 분포를 만들고자 하는 시도가 생겨났다.
  - 이 중 가장 유명한 모델로 GMM (Gaussian Mixture Model) 을 사용한다.



- 잠재 변수 (latent variable)을 도입하여 정교한 모델을 도입하게 된다.

K-means 알고리즘

- EM 과 K-means 알고리즘은 매우 관련이 깊다.

- 주어진 데이터를 K 개의 집합으로 클러스터링하는 문제이다.
  - 최종 목표는 주어진 데이터를 어느 하나의 클러스터에 속하도록 배정하는 것.
  - 이를 위해 K 개의 중심점(central point)을 지정하고,
  - 모든 데이터는 자신과 가장 가까운 중심점에 속하도록 학습한다.
  - 이를 위한 목적 함수는 다음과 같다.

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|x_n - \mu_k\|^2$$

Binary indicator variables

$$r_{nk} = \begin{cases} 1, & if\ k = argmin_j\|x_n - \mu_j\|^2 \\ 0, & otherwise \end{cases}$$

K-means 알고리즘

- MLE를 구해보자.

$$2\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \mu_k) = 0$$

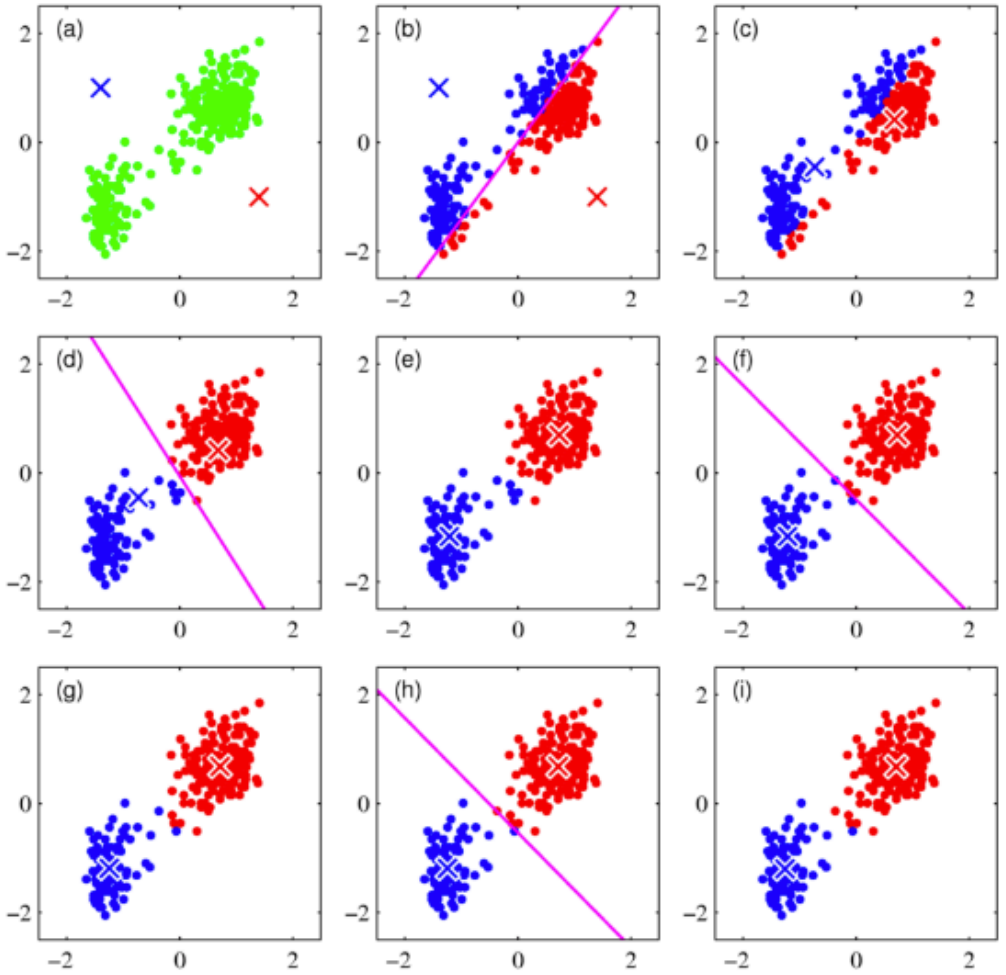
$$\mu_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$$

$$r_{nk} = \begin{cases} 1, & if\ k = argmin_j \|x_n - \mu_j\|^2 \\ 0, & otherwise \end{cases}$$

- 두 개의 파라미터가 서로 연관되어 있다.
- 두려워할 필요는 없음.
  - 각각 업데이트를 수행하는 방식을 채택
  - 반복하면서 수렴할때까지 진행
  - GD와 유사하다.

K-means 를 구하기 위한 EM

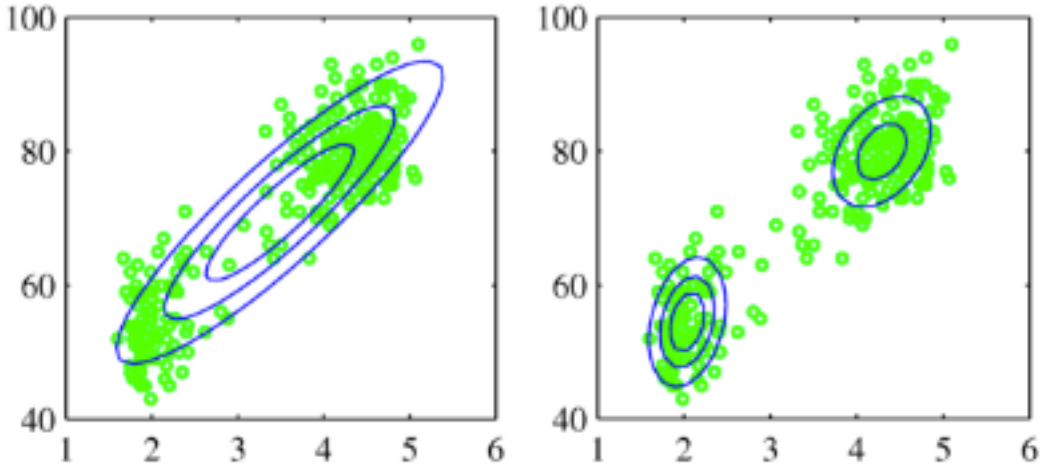
- r과 u 를 구하는 단계는 크게 2 단계로 나눔
  - 먼저 u를 임의의 값으로 초기화
  - u를 고정한 상태에서 J를 최소화하는 r을 구함.
  - r을 고정한 상태에서 u를 갱신
  - 적당히 수렴할 때까지 위 2단계를 반복
- 왼쪽 그림은 K=2 인 경우 문제 풀이
- 선을 긋는 과정이 r를 구하는 과정임.

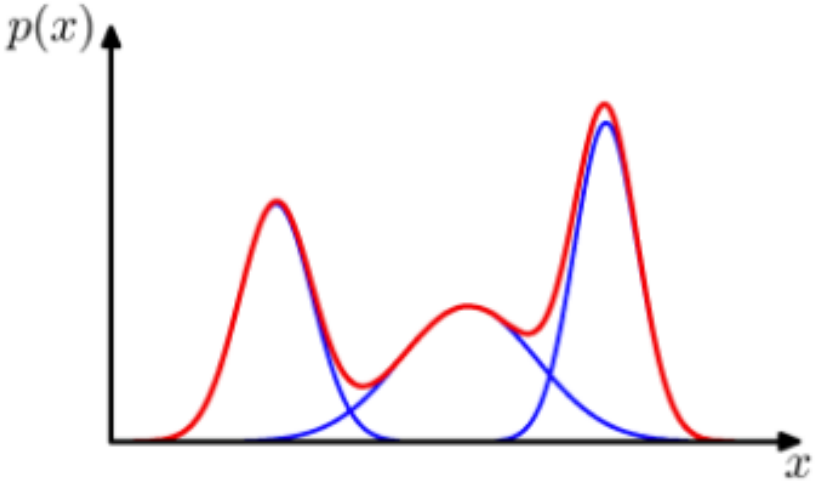


GMM (Gaussian Mixture Model)

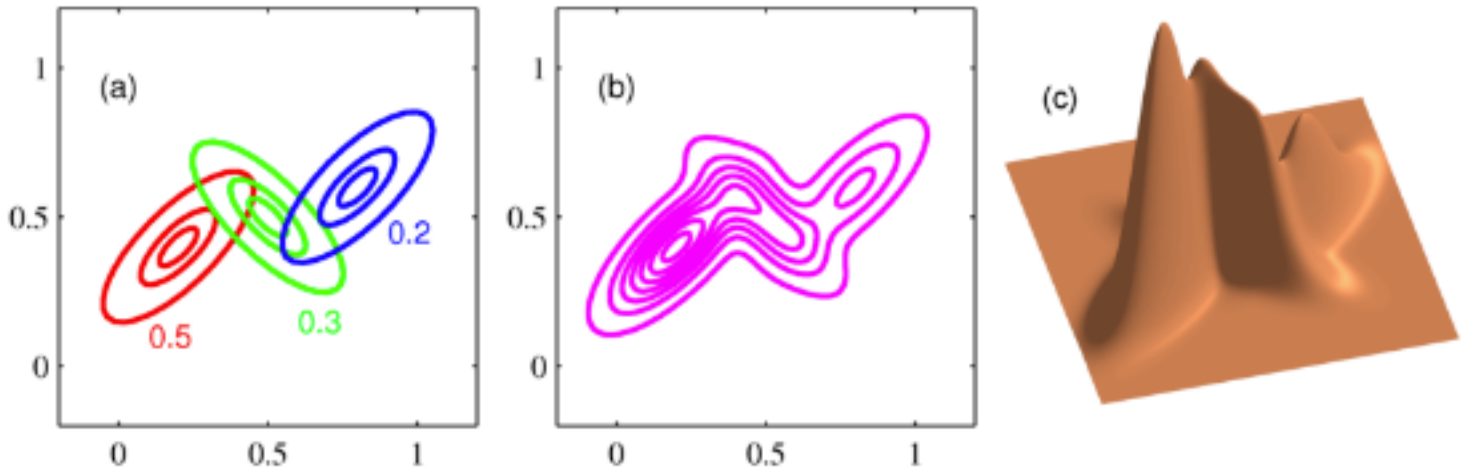
- 일단은 K-means 를 좀 더 다른 관점으로 확장한 것이라 생각해보자.
- 확률 분포 p(x)를 간단한 하나의 분포로 표현하기 어렵다는 문제의식
  - 이를 여러 개의 가우시안 분포의 선형 결합으로 다룬다는 아이디어.
  - K-means와 유사한 점은 하나의 데이터는 각각의 가우시안 분포에 특정 비율로 속하게 된다는 것.

$$p(x;\theta) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)$$





GMM (Gaussian Mixture Model) (Cont'd)



$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k N(\mathbf{x}|\mu_k, \Sigma_k)$$

$$\sum_{k=1}^{K} \pi_k = 1$$

$$0 \leq \pi_k \leq 1$$

GMM with latent variable

- 은닉 변수 z 를 도입하여 GMM 을 표현한다.

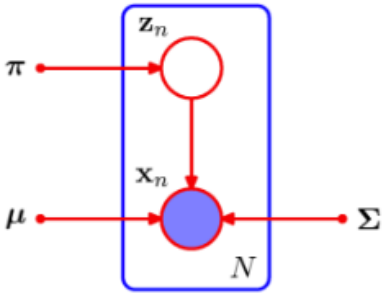
$$p(z_k = 1) = \pi_k \qquad p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \qquad p(\mathbf{x}|z_k = 1) = N(\mathbf{x}|\mu_\mathbf{k}, \Sigma_k)$$

$$p(\mathbf{x}) = \sum_{z} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k N(\mathbf{x}|\mu_k, \Sigma_k) \qquad p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$$

Responsibility

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_j^K p(z_j = 1)p(\mathbf{x}|z_j = 1)} = \frac{\pi_k N(\mathbf{x}|\mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j N(\mathbf{x}|\mu_j, \sigma_j)}$$

MLE for GMM



$$\ln p(\mathbf{X}|\pi,\mu,\Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(\mathbf{x}_n|\mu_k,\Sigma_k) \right\}$$

- 시간이 없으니 전개는 생략하고 바로 답을 적어본다.

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_j^K p(z_j = 1)p(\mathbf{x}|z_j = 1)} = \frac{\pi_k N(\mathbf{x}|\mu_k,\Sigma_k)}{\sum_{j=1}^K \pi_j N(\mathbf{x}|\mu_j,\sigma_j)}$$

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})\mathbf{x}_n \qquad \pi_k = \frac{N_k}{N} \qquad \Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$$

GMM 에 EM 적용하기.

- 초기화 단계 : 각 가우시안 분포의 평균, 공분산과 pi 값을 적당하게 초기화 한다.
- E 단계 : 주어진 파라미터 값을 이용하여 r 을 구한다.

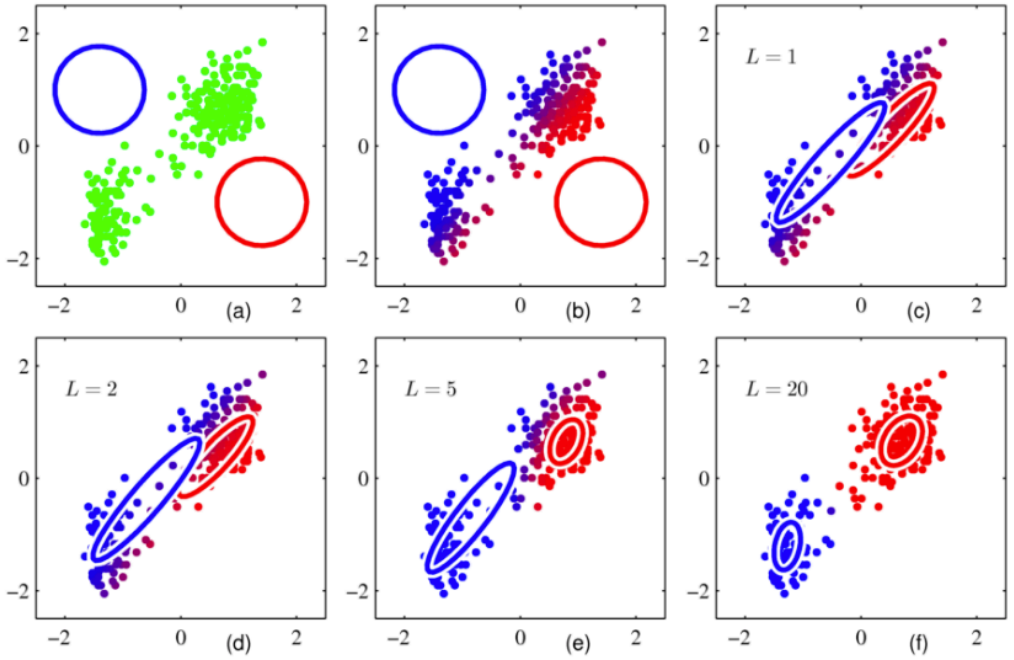
$$\gamma(z_{nk}) = \frac{\pi_k N(\mathbf{x}_n|\mu_k, \Sigma_k)}{\sum_j \pi_j N(\mathbf{x}_n|\mu_j, \Sigma_j)}$$

- M 단계 : 주어진 r 값을 이용하여 각각의 파라미터를 구한다.

$$\mu_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \qquad N_k = \sum_{n=1}^{N}\gamma(z_{nk})$$

$$\pi_k^{new} = \frac{N_k}{N} \qquad \Sigma_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$$

- 이를 반복한다.



EM 을 보는 또 다른 시선들.

- 앞서 지루하게 풀었던 문제들을 되돌아 보자.
- 모두 잠재 변수 z 를 도입하여 일반화된 모델로 만들 수 있다.
- z 에 대해 미리 알려진 바는 없다. 이를 추가하여 조건부 분포로 전개한다.
- 이런 방식이 EM 의 전형적인 방식.
  - SGD 와 다른 점은 z 를 도입하여 문제 풀이 방식을 고정한다는 것.
  - 생각보다 쉽게 모델을 설계할 수 있으므로 자주 사용하는 모델이다.
  - 하지만 발동할 수 있는 조건이 좀 까다롭다.

Expectation of log likelihood.

$$\ln p(\mathbf{X}|\theta) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta) \right\}$$

Not closed form

- 안타깝지만 위 식으로 기존의 MLE 방식을 적용할 수 없다.

- 대신 우리는 다음의 식을 최대화하는 방법으로 MLE를 푼다.
  - 주어진 샘플에 대해 아래 조건을 만족하는 “파라미터”와 “Z”를 구한다. (번갈아가며)

$$E_Z[\ln p(\mathbf{X}, \mathbf{Z}|\theta)] = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta) \ln p(\mathbf{X}, \mathbf{Z}|\theta)$$

General EM Algorithm

- 관찰데이터 $\mathbf{X}$ 와 잠재 변수 $\mathbf{Z}$ 가 파라미터 θ 에 의해 주어졌을 때,
- 결합 분포는 $p(\mathbf{X}, \mathbf{Z}|\theta)$ 와 같이 표현 가능하다.
- 이 때 $p(\mathbf{X}|\theta)$ 값을 가장 크게 만드는 파라미터 θ 값을 얻고 싶다. (MLE를 이용)
  - **Init Step** : (임의의) 파라미터 θ^{old} 의 값을 설정한다.
  - **E-Step** : $p(\mathbf{Z}|\mathbf{X}, \theta^{old})$ 값을 추론한다.
  - **M-Step** : θ^{new} 값을 추론한다. 이 때,
    - $\theta^{new} = \arg\max_{\theta} Q(\theta, \theta^{old})$ (9.32)
    - $Q(\theta, \theta^{old}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{old}) \ln p(\mathbf{X}, \mathbf{Z}|\theta)$ (9.33)
  - 새롭게 구해진 파라미터의 값들이 수렴 상태인지 확인한다.
    - 수렴되지 않았다면 아래 작업 후 Step-2 로 돌아간다.
    - 수렴되었을 경우 종료한다.
      - $\theta^{old} \leftarrow \theta^{new}$ (9.34)

EM 을 사용할 수 있는 경우란?

- 조건이 충족되어어야 EM 사용이 가능.
  - 일단 잠재 변수 z 를 도입할 수 있는 모델이어야 한다.
  - 잠재 변수 z 도입 전에는 문제 풀이에 어려움을 겪지만,
  - z 를 도입하고 나면 문제 풀이가 가능한 경우에만 사용할 수 있다.
    - 즉, 어떤 모델을 설계할 때 Z 가 알려지는 경우 우리가 잘 아는 분포로 도식화 가능하다고 생각되는 모델을 도입하게 된다. (GMM같은)

$$p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta)$$

EM 알고리즘 정리.

- 사실은 EM 알고리즘에 대한 충분한 이해가 선행되어야 한다.
  - MLE 문제도 좀 풀어보고 응용도 좀 풀어보고 해야 EM 스타일에 적응할 수 있다.

- 하지만 여기서 그런 것을 다룰 수는 없지 않는가!!! 다음만 기억하자.
  - 일단 z 를 알게되면 문제를 풀 수 있는 모델을 도입.
  - 하지만 실제로 z 를 알 방법은 없음.
  - 그래서 적당한 z를 추정하는 문제로 하여 반복하여 에러를 최소화하는 문제로 풀이

EM 알고리즘 정리. (Cont'd)

$$p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta)$$

- EM 은 잠재변수 Z를 가진 모델에서 MLE를 구하는 모델.
  - p(x) 는 일반적인 분포 형태가 아니어서 바로 추정이 어렵다. (incomplete-data)
  - p(x, z) 는 MLE로 쉽게 얻을 수 있는 분포. (complete-data)

- Z 는 보통 이산변수(즉, PMF)로 놓고 문제를 푼다.
  - 물론 실수 변수 (PDF)도 불가능한 것은 아니지만 구조를 머리로 그려보기 힘들다.

- 그럼 결국 p(z|x) 를 풀어야 하는 문제로 귀결됨.
  - 보통은 추정이 가능한 모델이라고 가정하고 문제를 푼다. (E-Step에서)
  - 그런데 그게 불가능하다면? -> VI

General EM Algorithm

q(z)의 도입

- 이제 VI 에서 무척이나 자주 등장하는 q(z)를 살펴보자.
- 정의는 그냥 너무 황당한데 앞서 보았던 z 와 관련된 그냥 어떤 함수 q() 다.
- 그냥 z와 관련된 어떤 함수가 q(z)가 존재한다고 할 때 앞선 식을 다음과 같이 전개 가능하다.

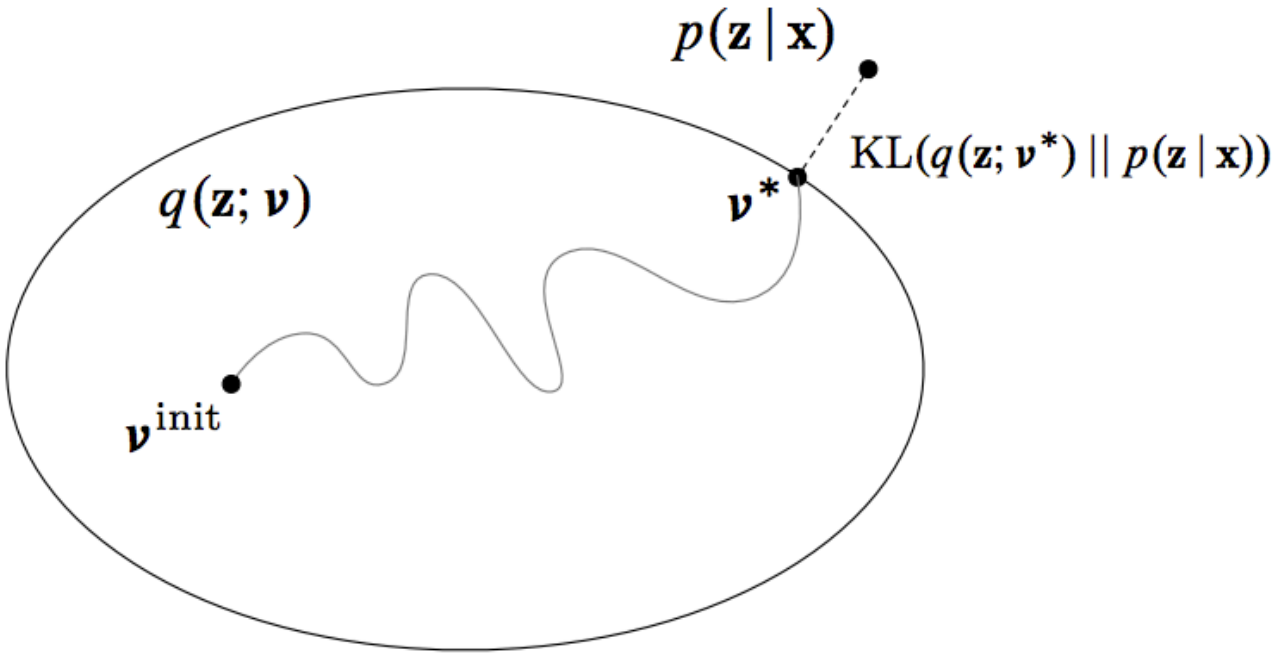
$$\ln p(\mathbf{X}|\theta) = L(q,\theta) + KL(q\|p)$$

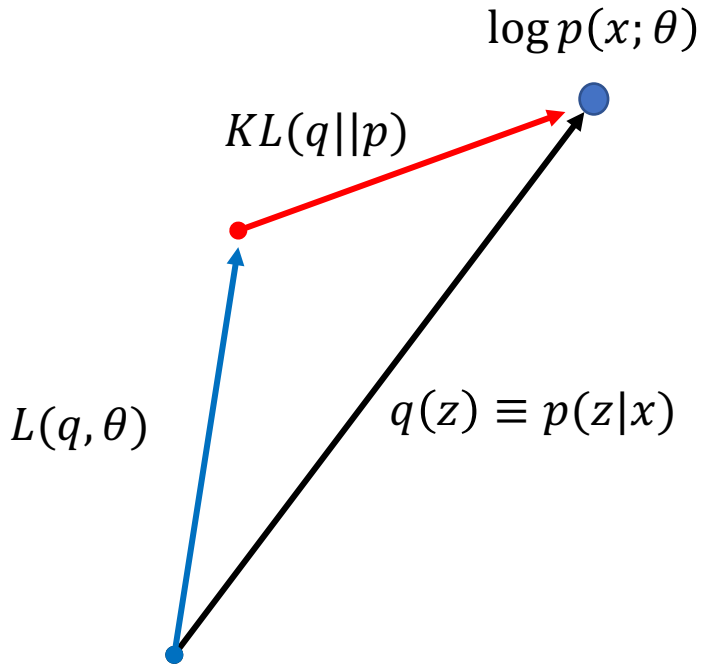
$$L(q,\theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln\left\{ \frac{p(\mathbf{X},\mathbf{Z}|\theta)}{q(\mathbf{Z})} \right\}$$

$$KL(q\|p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln\left\{ \frac{p(\mathbf{Z}|\mathbf{X},\theta)}{q(\mathbf{Z})} \right\}$$

$$p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\theta)$$

Variational Inference





EM for VI

$$\ln p(\mathbf{X}|\theta) = L(q,\theta) + KL(q\|p)$$

- L은 범함수이다. (입력인 q 가 함수임)
  - 하지만 θ 에 대해서는 그냥 함수.

$$L(q,\theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln\left\{ \frac{p(\mathbf{X},\mathbf{Z}|\theta)}{q(\mathbf{Z})} \right\}$$

- KL 은 앞서 살펴보았다.

$$KL(q\|p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln\left\{ \frac{p(\mathbf{Z}|\mathbf{X},\theta)}{q(\mathbf{Z})} \right\}$$

- 맨 처음 이 식이 등장하면 모두 멘붕됨.
  - 하지만 전개를 할 수 있다면 이미 용자.
  - 전개가 어렵더라도 개념만 알면 된다.

증명

- 증명을 생략하고 싶지만 아주 간단하게만 적어보자.
  - **Jensen's Inequality** (옌슨 부등식) - 함수 f 가 convex 인 경우 다음을 만족.

$$E[f(x)] \geq f(E[x])$$

참고로 concave 인 경우 반대 성립

  - **Gibb's Inequality** (깁스 부등식) - p 와 q 에 대해 항상 다음을 만족한다.

$$KL[p\|q] \geq 0$$

  - 단, p==q 인 경우 0, 아닌 경우 0 보다 크다.

증명 (Cont'd)

$$\ln \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta) = \ln \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right\} = L(q, \theta)$$

$$\ln p(\mathbf{X}|\theta) - L(q, \theta) = \ln p(\mathbf{X}|\theta) - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\}$$

$$= \ln p(\mathbf{X}|\theta) - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \theta) p(\mathbf{X}|\theta)}{q(\mathbf{Z})} \right\}$$

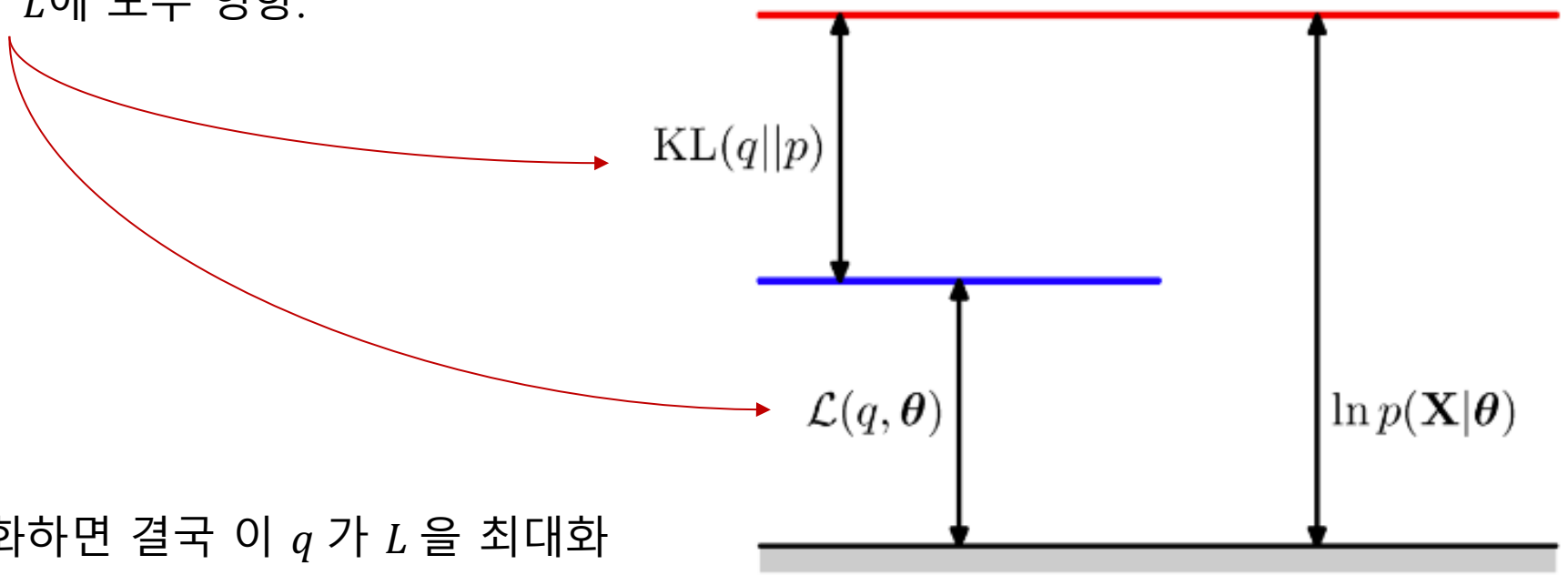
$$= \ln p(\mathbf{X}|\theta) - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})} \right\} - \ln p(\mathbf{X}|\theta) \sum_{\mathbf{Z}} q(\mathbf{Z})$$

$$= -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})} \right\} = KL[q(\mathbf{Z}) \| p(\mathbf{Z}|\mathbf{X}, \theta)] = KL[q \| p]$$

Decomposition of $p(x)$

- $q = p$ 인 경우 $KL(q||p) = 0$ 이 된다.

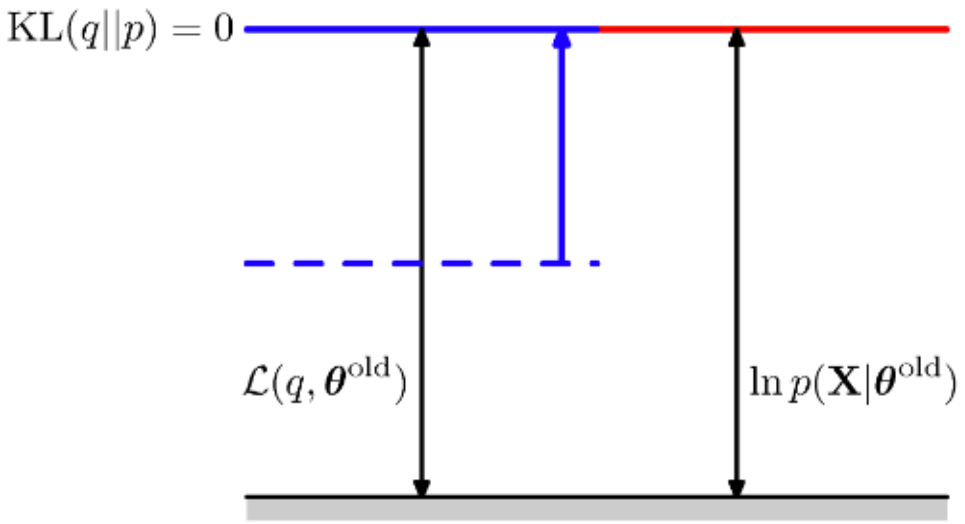
이 때 q 는 KL과 L에 모두 영향.



$KL(q||p)$를 최소화하면 결국 이 q 가 L 을 최대화

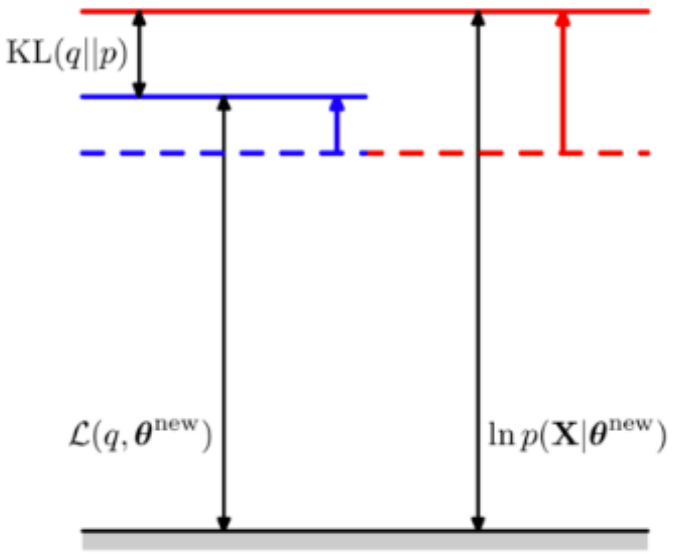
서로 상보적 관계가 된다.

E-Step



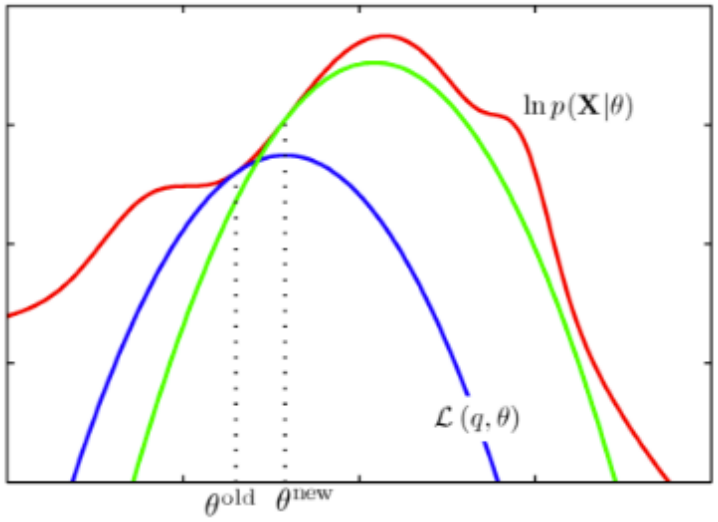
- 파라미터 θ 를 고정시킨 상태에서 함수 q 를 최적화
- $KL(q||p) = 0$ 상태를 만드려고 노력한다.
- 적합한 q 를 구할수록 Bound 값인 L 은 커진다.

M-Step



- 함수 q 를 고정한 상태에서 파라미터 θ 를 최적화
- 보통 MLE 를 활용하여 최적의 θ 를 찾게 된다.
- 새로 생성된 θ 에 의해 Bound 값인 L 은 커진다.

Overview for EM



- 이 그림이 EM의 모든 것을 요약하는 그림이다.
- 우선 임의의 파라미터 θ_{old} 로부터 EM 알고리즘이 시작됨
- 해당 지점에서 로그 가능도 함수와 최대한 근사한 L 함수를 만들게 된다.
  - 이 때 L 함수는 θ_{old} 에 대해 concave 함수이으로 위와 같은 그림이 된다. (파란색)
  - 이 단계가 E-Step에서 이루어진다.
- 얻어진 L 함수를 최대화하는 새로운 파라미터 값을 선정하게 된다.
  - 위의 그림에서는 θ_{new} 이다.
  - 이를 이용하여 새로운 L 함수를 얻음. (초록색)
  - 이 단계가 M-Step 단계이다.
- 수렴 조건을 만족할 때까지 반복하게 된다.

Variational Inference

이제 진짜 시작입니다.

변분법 (Variational Method)

- 변분법은 오일러와 라그랑지안의 변분 이론에서 출발
  - 보통 함수는 실수를 입력받아 실수를 반환하게 되는데,
  - 함수를 입력으로 받는 함수도 생각할 수 있다. : 이게 바로 범함수(functional).
  - 범함수를 함수로 미분하는 문제.
    - 입력으로 사용되는 함수를 조금 바꿀 때 함수의 출력값의 변화량을 측정.

- 변분법 자체는 근사 기법이 아니다.
  - 하지만 VI 에서는 이런 함수를 고정된 형태로 제한하여 결국엔 함수를 근사하도록 함.
  - 제한하는 함수 형태란?
    - 이차함수(quadratic) or 기저 함수를 포함한 선형 결합 or 인수 분해

VI 모델의 가정

- VI 는 추론에 있어 모든 파라미터가 사전 분포를 가진다고 가정.
  - 즉, Full Bayesian 모델이 된다.
  - 그리고 이러한 모든 파라미터를 변수 Z 에 포함한다.
  - 따라서 사용되는 모든 파라미터 또한 latent variable 로 간주된다.
  - 일단 모든 latent variable 은 연속형 변수(continuous random variable)로 취급한다.
    - 이후 이산 변수로의 변환이 필요하면 적절하게 적분을 합산으로 변경하면 된다.

- 목표
  - 모델 $p(x)$ 와 사후 분포 $p(z|x)$ 를 위한 최적의 근사 함수를 찾는 것.

베이지언 모델로의 통합.

$$\ln p(\mathbf{X}) = L(q) + KL(q\|p)$$

$$L(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} d\mathbf{Z}$$

$$KL(q\|p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} d\mathbf{Z}$$

- 위 식을 보면 사용되던 파라미터들을 모두 잠재 변수로 통합되었다.
  - 앞서 설명한 대로 모든 파라미터를 랜덤 변수로 취급하는 Full 베이지언 모델을 채택

- 기본적인 식은 연속형 변수로 취급
  - 필요한 경우 적분을 합산 공식으로 변환하여 사용하면 된다.

베이지언 모델로의 통합. (Cont'd)

- 모든 Z 에 대해 제한적인 계열(family) 의 분포를 가정.
- KL을 최소화하기 위한 파라미터 값을 추정
- 근사 분포 정하기
  - 파라미터를 사용하는 분포 도입 : $q(z; w)$
  - 분포 계열 제한.
  - Factorized distributions

Factorized distributions

$$q(\mathbf{Z}) = \prod_{i}^{M} q_i(\mathbf{Z}_i)$$

- 각각의 z 는 적당한 단위의 q 함수로 나누어질 수 있다.
- 즉, 임의의 z 에 대해 독립적이라는 가정을 할 수 있음.
- 이런 가정 하에서 $L(q)$ 를 최대화하는 q 를 구한다.
- 이러한 기법을 평균장 이론 (mean-field theory) 이라고 한다.

평균장 이론

- 평균장 이론물리학에서 개발된 근사 프레임워크로 자기 모순 없는 장 이론.
  - (self-consistent field theory)라고도 함
- 다수의 상호작용이 있는 복잡한(many-body) 문제를

  단순한 하나의 상호작용(one-body)의 단순 모델로 표현하는 방법

- 각각의 요소에 대한 상호작용을 이해하고 계산하기 어려우니, 평균 상호 작용으로 취급하는 것

Factorized distributions (Cont'd)

$$L(q) = \int \prod_i q_i \times \left\{ \ln p(\mathbf{X}, \mathbf{Z}) - \sum_i \ln q_i \right\} d\mathbf{Z}$$

$$= \int q_j \prod_{i \neq j} q_i (\ln p(\mathbf{X}, \mathbf{Z}) - \ln q_j) d\mathbf{Z} - \int q_j \prod_{i \neq j} q_i \sum_{i \neq j} \ln q_i d\mathbf{Z}$$

$$= \int q_j \left\{ \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i d\mathbf{Z}_i \right\} d\mathbf{Z}_j - \int q_j \ln q_j d\mathbf{Z}_j + const$$

$$= \int q_j \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) d\mathbf{Z}_j - \int q_j \ln q_j d\mathbf{Z}_j + const$$