▶목적/손실 함수(Loss Function)은 무엇일까?

딥러닝과 머신러닝은 결국 컴퓨터가 어떤 해를 찾아가는 과정이다.

Loss Function은 현재값이 실제값과 가까워지게 그 차를 찾아가는 함수를 말한다.

간단한 문제를 통해 이해해보자

문제1

$$3=2a+b$$
$$1=a+b$$

문제1을 해결하기 위해 가능한 솔루션은

1. 값을 직접 대입해본다.
2. 직관적으로 풀어본다.
3. 여러가지 공식을 써서 풀어본다.

정도일 것이다.

여기서 2번과 3번의 경우에는 컴퓨터에게 시키기 매우 힘든 작업이다.

문제1
3=2a+b
1=a+b

예를 들어 사람이라면 어떻게 풀까?

2번식에서 1번식을 빼면 a=2라는 값을 쉽게 얻을 수 있다. 그를 통해 2번식에 대입하여 b=-1이라는 결과를 낼 것이다. 그러나 비슷한 문제에서 컴퓨터가 이러한 풀이법을 코드로 구현하는 것은 인간처럼 직관적이지 않다.

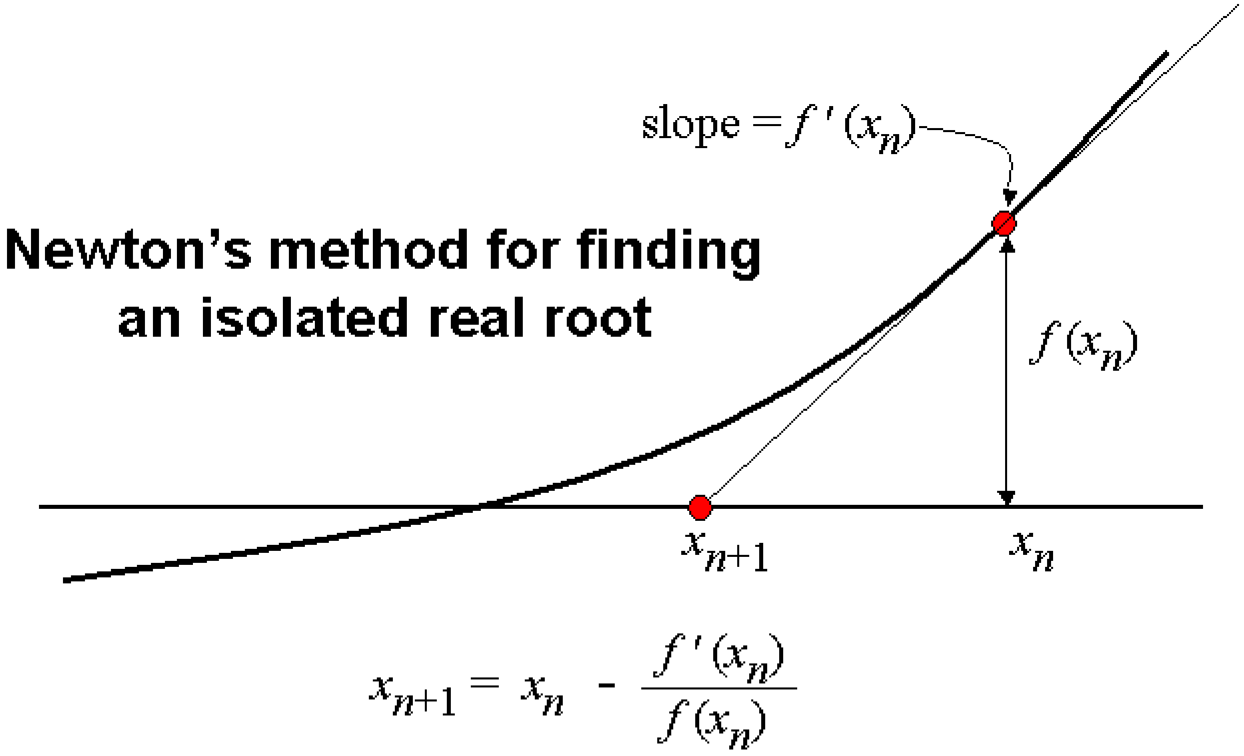
(물론 예제는 매우 쉬운 문제이기에 어렵지 않다)

그렇기 때문에 컴퓨터는 보통 1번의 방법(값을 직접 대입)을 채택한다.

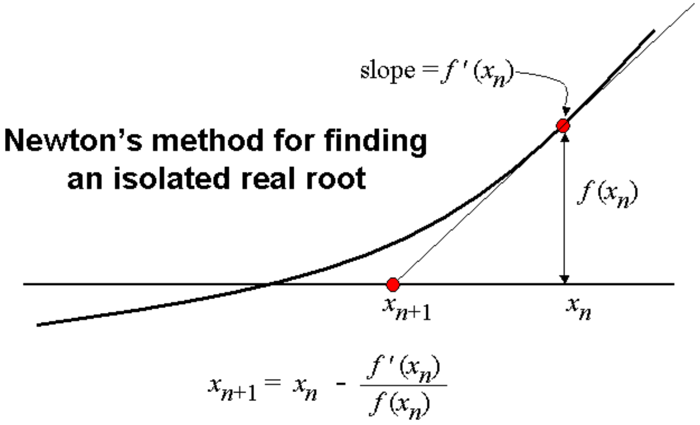
PPT PRESENTATION

쉽게 Loss function의 예시로는 가장 기본적인 모델 중 하나인

Newton’s method 등이 있다.



PPT PRESENTATION



간단하게 y= x^2-2의 해(1.414...)를 구한다 생각하자 도함수는 2x이다

컴퓨터는 x에 적당한 1을 대입한다.(x0)

그럼 y의 값은 -1, 도함수 값은 2가 나올 것이다. 이 값을 바탕으로

X1=X0-(-1)/2=1.5로 1.414에 1보다 근접한 값이 나온다. 이것을 반복하

여 해를 구하는 것이다.

PPT PRESENTATION

▶ Notation

$\mathcal{X}$: a compact metric set, x가 가질 수 있는 값들을 모아놓은 집합이다.
$\mathbb{P}$: 분포 또는 확률 measure이다. $\mathbb{P}(x)$라 하면 분포 $\mathbb{P}$에서 x가 등장할 확률이 된다.
P : pdf(Probability Density Function)이다. $P(x)$가 확률이 아니라 $\int P(x)dx$가 확률이다!! ($\int Pd\mu$)
$Supp(f)$: support라고 한다. 함수(mapping) f가 0이 되지 않게 하는 f의 정의역의 부분집합이다. 예시로 ReLU는 실수 전체가 정의역이지만 양수일 때만 0이 아닌 값을 가지므로 $Supp(ReLU) = \mathbb{R}^+$가 된다.

출처
https://haawron.tistory.com/21

PPT PRESENTATION

Unsupervised learning의 대표적인 예제인 Maximum Likelihood Estimation(MLE)은 아래의 식을 푸는 것으로 나타낼 수 있다.

θ로 parameterized된 분포들의 모임 $(P_\theta)_{\theta \in \mathbb{R}^d}$이 있을 때, 실제로 수집한 데이터 샘플 $\{x^{(i)}\}_{i=1}^m$에 대하여

$$\max_{\theta \in \mathbb{R}^d} \frac{1}{m} \sum_{i=1}^{m} \log P_\theta(x^{(i)})$$

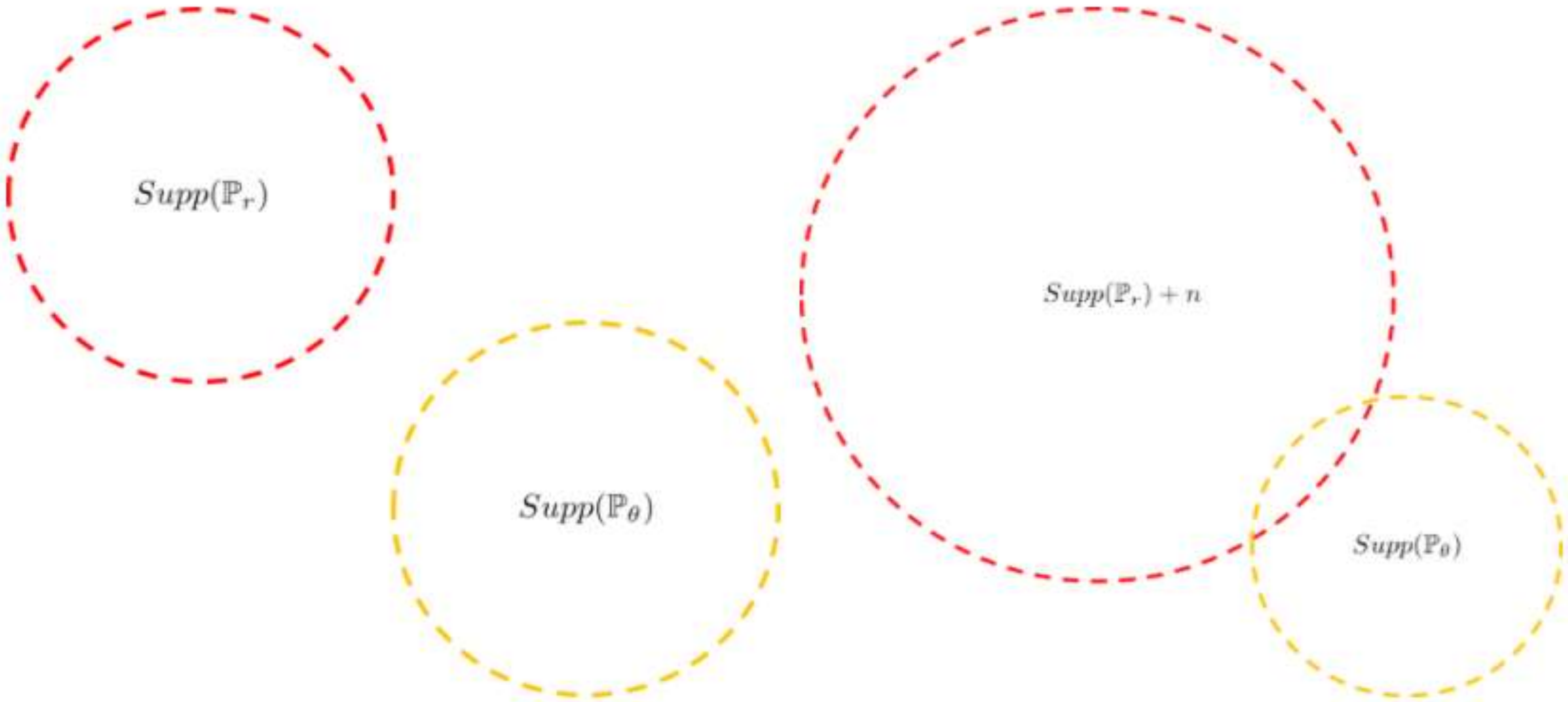
(연속 확률 분포에서는 likelihood를 measure $\mathbb{P}$가 아니라 density P로 정의한다.)

식의 의미를 표현하자면 "Data 덩어리 $\mathcal{X}$를 가장 잘 표현하는 분포 $\mathbb{P}_\theta$를 찾아라!" 정도 되겠다. 이 문제를 푸는 건 KL-divergence를 최소화하는 것과 동치라고 한다.

PPT PRESENTATION

KL-divergence는 분포들의 집합 $\{\mathbb{P}\}$에서 정의된 하나의 거리와 같다. MLE를 잘 풀려면 그의 동치인 KL-divergence가 0으로 잘 수렴해주어야 하는데, 애초에 두 분포의 support가 겹치지 않으면 KL이 발산하기 때문에 계산조차 불가능하다. 우리가 풀고자 하는 이미지 생성 문제는 이미지들이 아주 고차원에 분포해있기 때문에 이런 문제들이 더 빈번하게 발생한다.

Support가 강제로 겹치게 해줄 수 있다. 기존 이미지에 가우시안 등의 노이즈를 추가해주면 아래 그림처럼 조금이나마 겹치게 할 수 있다. 하지만 그렇게 좋은 해법은 아니고 생성된 이미지도 굉장히 흐릿하다고 한다.

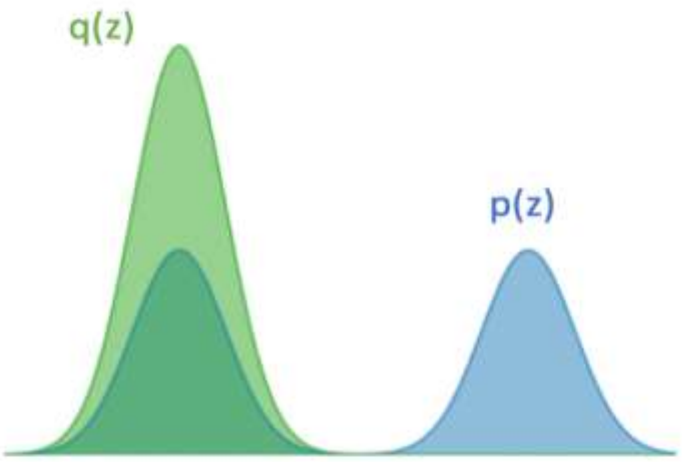


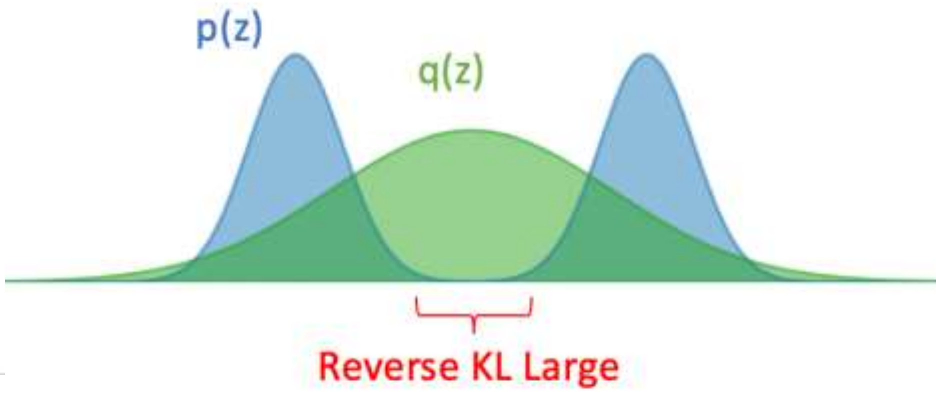
The *Kullback-Leibler* (KL) Divergence

일단 KLD는 다음과 같이 정의된다.

$$KL(\mathbb{P}_r || \mathbb{P}_g) = \int \log\left(\frac{P_r(x)}{P_g(x)}\right) P_r(x) d\mu(x)$$

식을 봤을 때 가장 불안한 요소라 하면 분모에 있는 P_g이다. 이 때문에 $P_g(x) = 0$이지만 $P_r(x) \neq 0$인 곳이 생긴다면 발산하게 된다. 논문에서는 이런 일이 저차원에서 빈번하게 일어난다고 한다.





The *Jensen-Shannon* (JS) Divergence

JS는 KL을 이용해 간단하게 표현할 수 있다.

$$JS(\mathbb{P}_r||\mathbb{P}_g) = \frac{1}{2}KL(\mathbb{P}_r||\mathbb{P}_m) + \frac{1}{2}KL(\mathbb{P}_g||\mathbb{P}_m), \text{ where } \mathbb{P}_m = (\mathbb{P}_r + \mathbb{P}_g)/2$$

(논문에 1/2이 빠졌다.)

$\mathbb{P}_m = 0$이면 $\mathbb{P}_r = \mathbb{P}_g = 0$이기 때문에 발산할 일은 없다.

$$\because \mathbb{P}_m = 0 \Longrightarrow \mathbb{P}_r = -\mathbb{P}_g \geq 0, \text{ and } \mathbb{P}_r, \mathbb{P}_g \geq 0$$

PPT PRESENTATION

하지만 두 분포의 support가 겹치지 않는다면

$$P_g(x) \neq 0 \Rightarrow P_r(x) = 0$$
$$P_r(x) \neq 0 \Rightarrow P_g(x) = 0$$

이기 때문에

$$\begin{aligned}
JS(\mathbb{P}_r||\mathbb{P}_g) &= \frac{1}{2}\left(\int_{\mathcal{X}} \log\left(\frac{P_r(x)}{(P_r(x)+P_g(x))/2}\right)P_r(x)d\mu(x) + \int_{\mathcal{X}} \log\left(\frac{P_g(x)}{(P_r(x)+P_g(x))/2}\right)P_g(x)d\mu(x)\right) \\
&= \frac{1}{2}\left(\int_{Supp(P_r)} \log\left(\frac{P_r(x)}{P_r(x)/2}\right)P_r(x)d\mu(x) + \int_{Supp(P_g)} \log\left(\frac{P_g(x)}{P_g(x)/2}\right)P_g(x)d\mu(x)\right) \\
&= \frac{\log 2}{2}\left(\int_{Supp(P_r)} P_r(x)d\mu(x) + \int_{Supp(P_g)} P_g(x)d\mu(x)\right) \\
&= \log 2
\end{aligned}$$

The *Earth Mover* (EM) Distance or *Wasserstein-1*

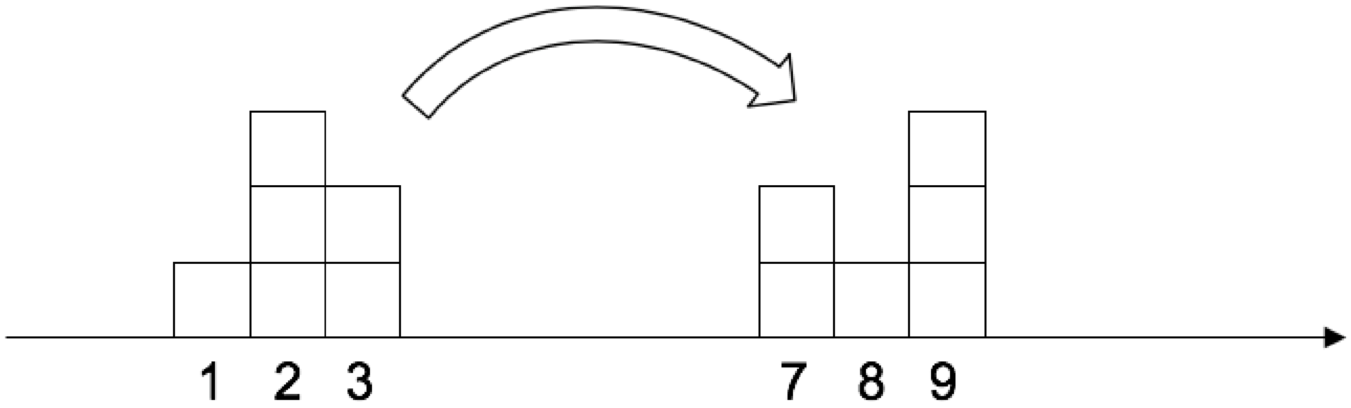
논문에서 소개하는 distance이다.

γ : 를 $\mathbb{P}_r, \mathbb{P}_g$간의 joint distribution 중 하나 (= coupling)
$\Pi(\mathbb{P}_r, \mathbb{P}_g)$: marginal이 $\mathbb{P}_r, \mathbb{P}_g$인 모든 joint distribution들의 집합
이라 했을 때 다음과 같이 정의된다.

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma} [||x - y||]$$

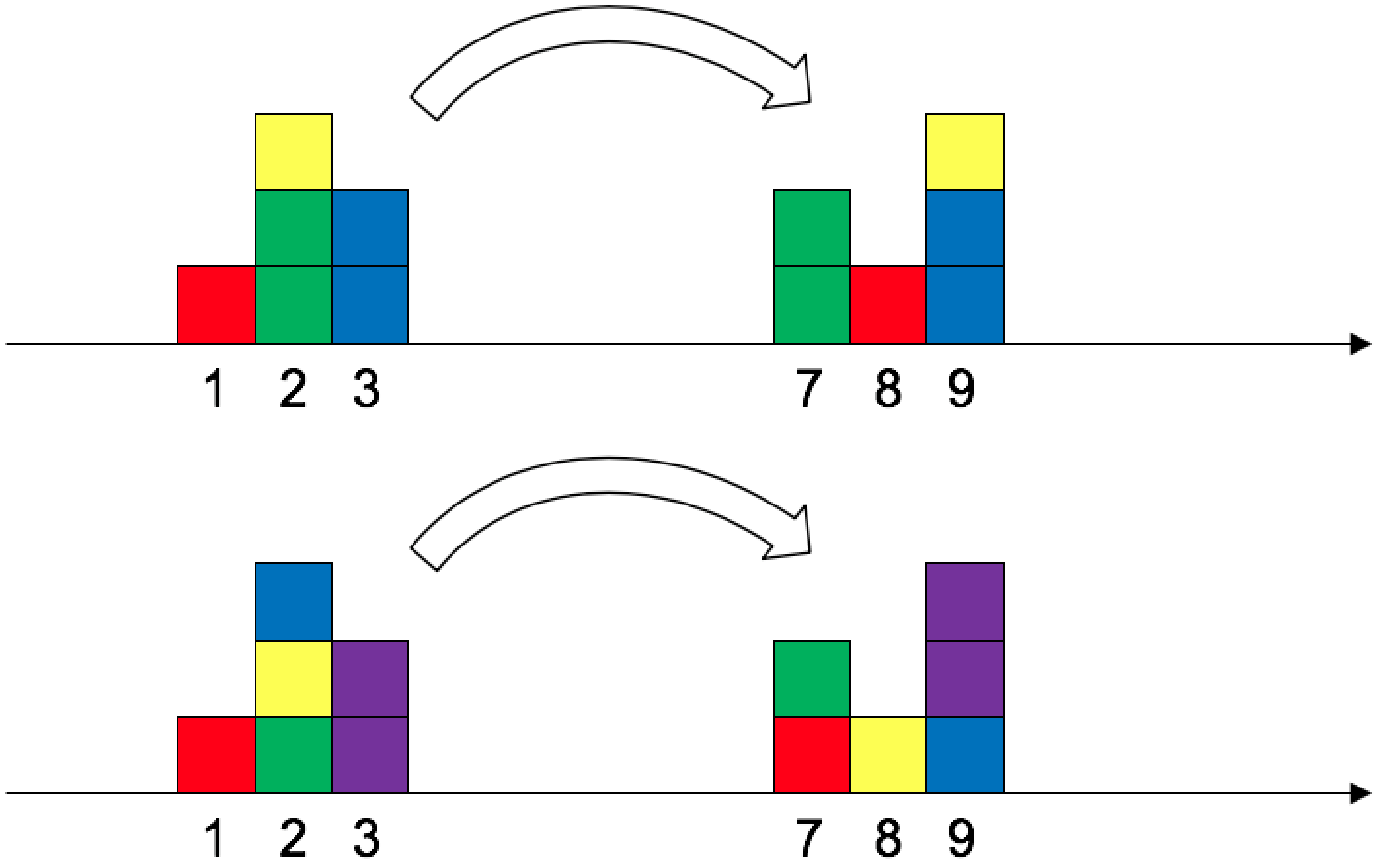
PPT PRESENTATION

Earth-Mover란 흙을 파서 옮기는 기계라고 한다. 여기서의 의미도 크게 다르지 않다. 흙을 파서 옮기데에 드는 비용을 distance로 표현한 것이라 할 수 있겠다.



PPT PRESENTATION

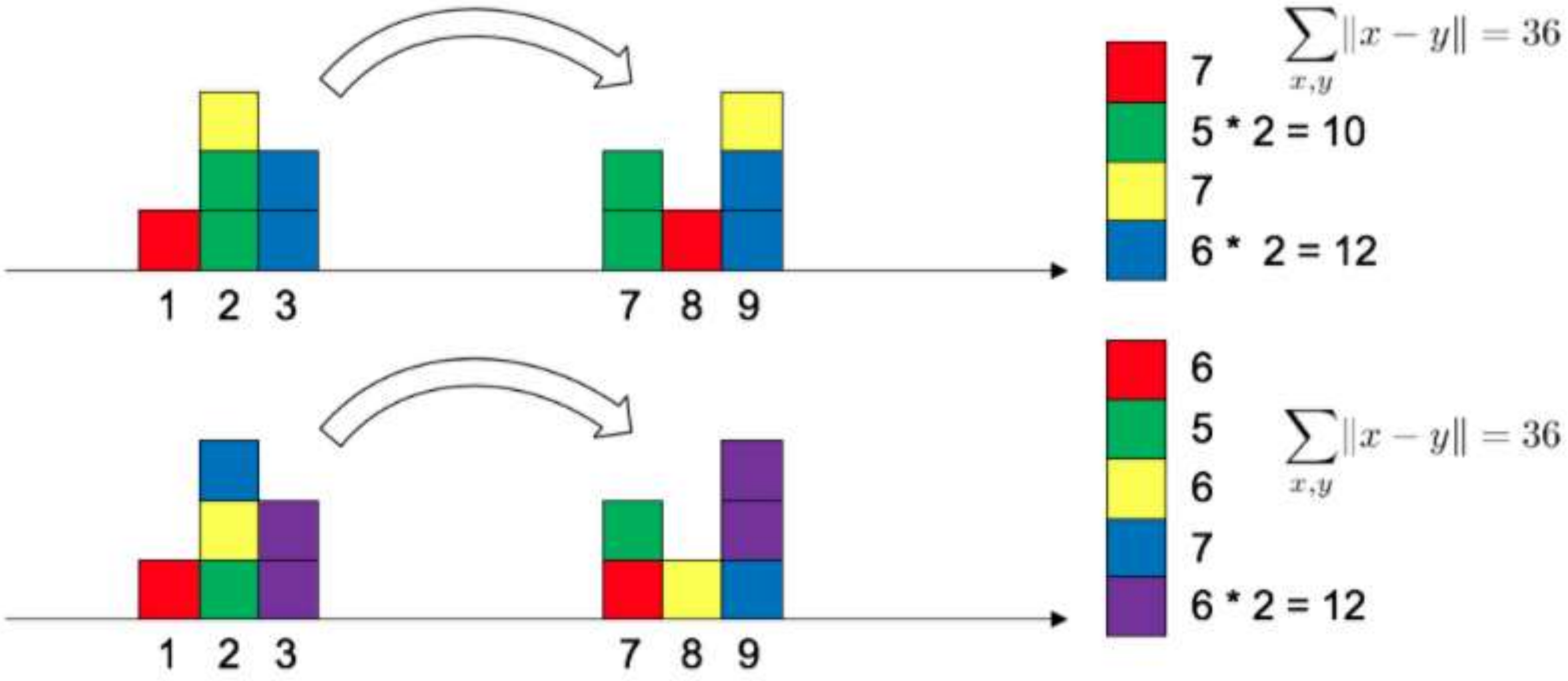
우리는 확률분포를 블록이 쌓여 있는 것으로 생각할 수 있다. 왼쪽은 $\mathbb{P}_r$, 오른쪽은 $\mathbb{P}_g$를 나타낸 것이고 왼쪽에 쌓인 8개 블록을 오른쪽 모양과 같이 옮길 것이다. 수 많은 방법 중 두 가지를 그림으로 표현해봤다.



그리고 각각의 방법의 transportation cost를 계산해볼 수 있다.

PPT PRESENTATION

그리고 각각의 방법의 transportation cost를 계산해볼 수 있다.

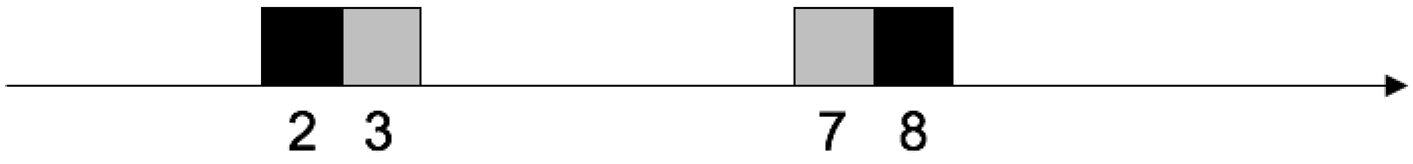


높이는 생각하지 말자

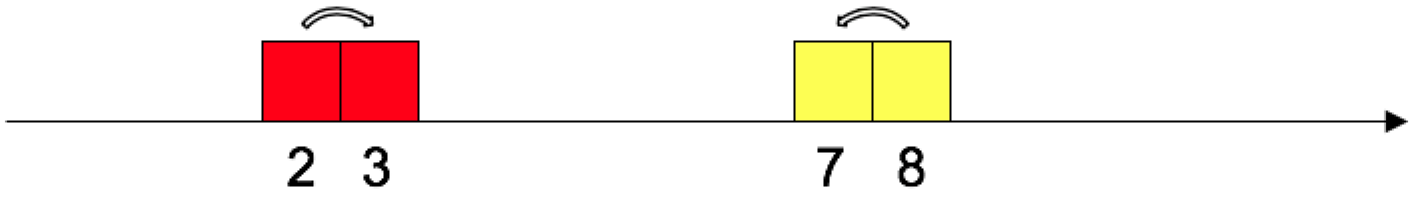
두 방법의 cost가 같은 것을 확인했다!

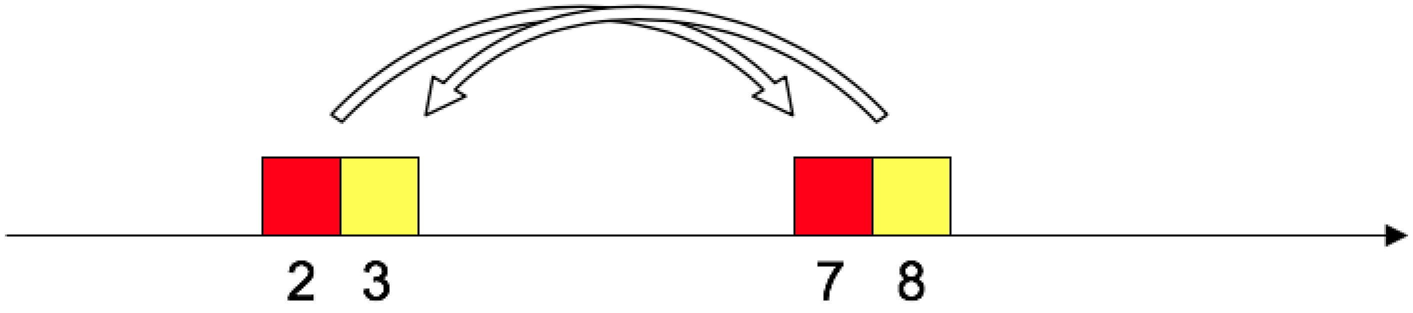
PPT PRESENTATION

하지만 항상 이렇게 cost가 같은 상황만 있는 것은 아니다.



위 그림에서 검은색 블록을 회색 블록이 있는 곳으로 옮긴다고 생각해보자. 그러면 다음의 두 가지 방법이 있을 것이다.

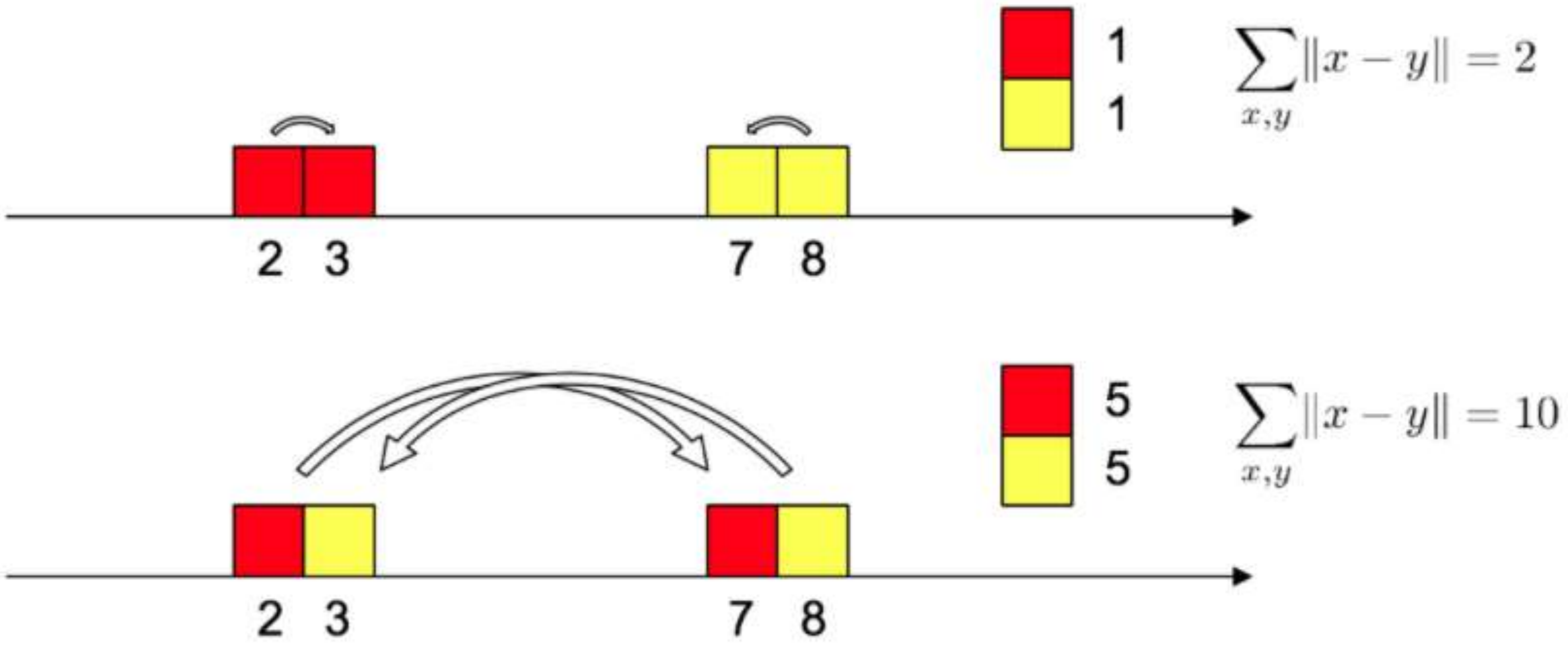




위쪽 방법이 더 경제적이므로 cost가 낮아야 한다!

PPT PRESENTATION

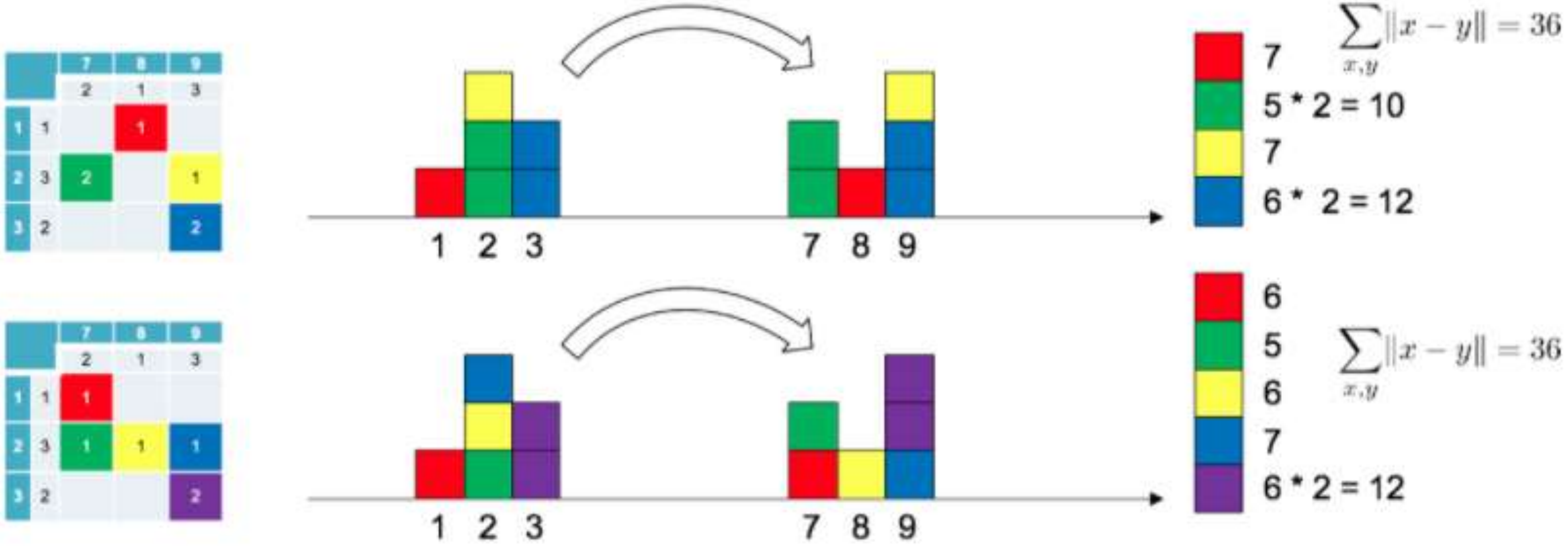
블록을 바로 옆으로 한 칸씩 옮기면 best겠지만 아래쪽처럼 멀리 옮길 수도 있겠다. 이를 transportation cost로 표현하면



2의 힘만 들이고도 목표한 위치로 옮길 수 있다는 것을 알 수 있다. 그리고 지금은 옮기는 방법이 두 개 밖에 없는 것이 확실하므로 cost의 최솟값이 2인 것도 확실하다.

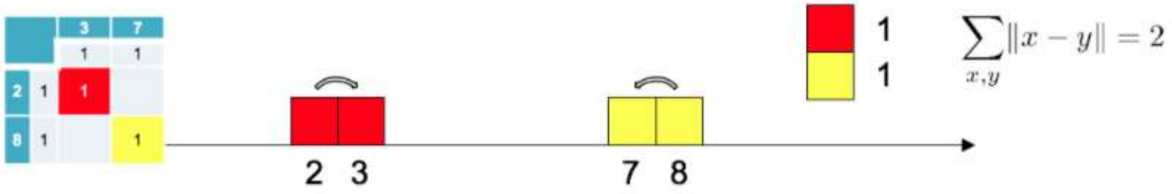
PPT PRESENTATION

사실 우리가 위에서 생각해 본 하나 하나의 plan을 joint probability distribution으로 생각할 수 있다.



표의 세로축이 1, 2, 3에 있는 블록 개수, 가로축이 7, 8, 9에 있는 블록 개수를 뜻한다. 왼쪽 표의 모든 값들을 6으로 나눠주면 모든 요소의 합이 1이므로 joint probability distribution으로 생각할 수 있다.

PPT PRESENTATION



위의 경우도 joint probability distribution으로 생각하면

		3	7
		1	1
2	1	1	
8	1		1

		3	7
		0.5	0.5
2	0.5	0.5	
8	0.5		0.5

왼쪽은 그냥 블록 옮기기 plan, 오른쪽은 이를 2(블록 개수)로 나눠 probability distribution으로 표현한 것이다.

위 그림의 오른쪽 표 같이 나타낼 수 있는데 확률로 나타냈으니 이제 W를 계산할 수 있다.

$$W(\mathbb{P}_r, \mathbb{P}_g) = \gamma_{X,Y}(2,3) \times |2-3| + \gamma_{X,Y}(8,7) \times |8-7| = 0.5 \times 1 + 0.5 \times 1 = 1$$

이를 continuous한 경우로 확장시키면 금방 이해 될 것이다.

4주차 end

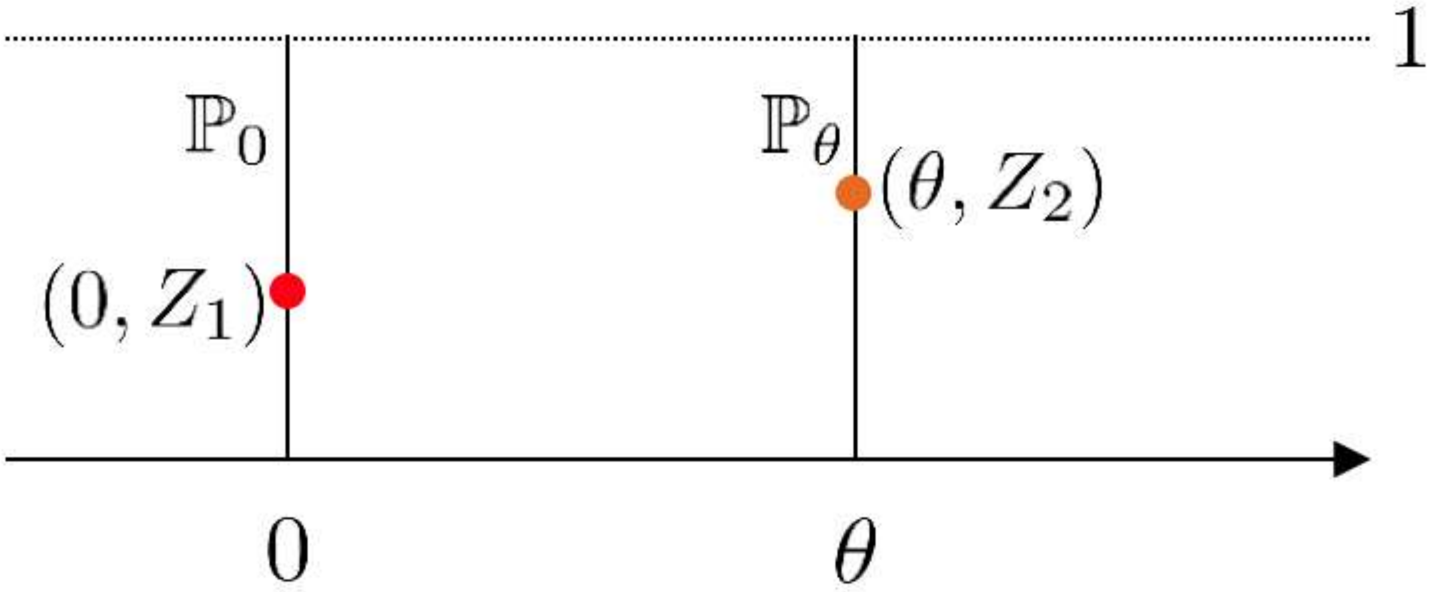
PPT PRESENTATION

PPT PRESENTATION

PPT PRESENTATION

$\mathbb{P}_0$는 x좌표는 $0, y$좌표는 $0 \sim 1$ 사이의 값을 가지는 점들의 분포이고,
$\mathbb{P}_\theta$는 x좌표는 θ, y좌표는 $0 \sim 1$ 사이의 값을 가지는 점들의 분포이다.

그림으로 보면 그렇게 어려운 내용은 아니다.



Wasserstein GAN

W distance가 GAN의 loss로 써먹기에 좋다는 것을 알아내긴 했지만

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}[||x-y||]$$

수식의 계산이 힘들다. $(x, y) \sim \gamma$는 sampling으로 해결이 가능하지만 Π는 space가 워낙 넓어 탐색하기도 힘들고 최솟값에 대한 보장도 없을 것이다. 그래서 이 논문에서 새로운 trick을 제안한다.

Kantorovich-Rubinstein Duality Theorem 라는 것을 이용하면

$$W(\mathbb{P}_r, \mathbb{P}_g) = \sup_{||f||_L \le 1} \mathbb{E}_{x\sim\mathbb{P}_r}[f(x)] - \mathbb{E}_{x\sim\mathbb{P}_\theta}[f(x)]$$

가 된다! $||f||_L \leq 1$은 f가 1-립쉬츠 함수(임의의 두 점 사이의 평균변화율이 1을 넘지 않는 함수)라는 뜻이다.

$$\frac{\Delta y}{\Delta x} = \frac{f(b)-f(a)}{b-a} = \frac{f(a+\Delta x)-f(a)}{\Delta x}$$