chapter 18
Confronting the Partition Function

인공지능학과 김다영

들어가기 전에,

비정규화 확률분포 $\tilde{p}(\mathbf{x};\theta)$ 로 정의되는 것들이 많음

-> 무향 그래프 모형

무향 그래프 모형을 사용할 때에는 비정규화 확률분포 $\tilde{p}$
분배함수 Z(θ) 로 나누어서 유효한 확률분포를 구해야 함

들어가기 전에,

정규화 상수로 쓰이는 분배함수는 모든 상태의 비정규화 확률에 대한 적분 또는 합

$$p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})}\tilde{p}(\mathbf{x};\boldsymbol{\theta})$$

연속 변수인 경우

$$\int \tilde{p}(\boldsymbol{x})d\boldsymbol{x}$$

이산변수인 경우

-> 적분이나 합을 구하는 연산은 처리 불가능

18.1 로그가능도의 기울기

$$\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x};\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log \tilde{p}(\mathbf{x};\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} \log Z(\boldsymbol{\theta}).$$

-> 양의 단계와 음의 단계로 분해하는 방법

-> 대부분의 무향 모형에서는 음의 단계를 계산하기 어려움

-> 잠재변수가 없거나 잠재변수들 사이의 상호작용이 적은 모형에서는 양의 단계가 처리 가능한 수준일 때가 많음

18.1 로그가능도의 기울기

양의단계는 쉽지만 음의단계가 어려운 모형의 좋은 예

-> RBM

Restricted Boltzmann Machine

visible layer(가시 층)와 hidden layer(은닉 층)으로 구성된 이진 확률적 그래픽 모델

-> 은닉 단위들은 가시 단위들이 주어졌을 때 서로 조건부 독립이기 때문에
양의 단계가 더 쉬움

18.1 로그가능도의 기울기

logZ의 기울기

$$\nabla_{\boldsymbol{\theta}} \log Z = \frac{\nabla_{\boldsymbol{\theta}} Z}{Z}$$

$$= \frac{\nabla_{\boldsymbol{\theta}} \sum_{\mathbf{x}} \tilde{p}(\mathbf{x})}{Z}$$

$$= \frac{\sum_{\mathbf{x}} \nabla_{\boldsymbol{\theta}} \tilde{p}(\mathbf{x})}{Z}.$$

$\tilde{p}(\mathbf{x})$ 에 $\exp(\log \tilde{p}(\mathbf{x}))$

$$\frac{\sum_{\mathbf{x}} \nabla_{\boldsymbol{\theta}} \exp(\log \tilde{p}(\mathbf{x}))}{Z}$$

$$= \frac{\sum_{\mathbf{x}} \exp(\log \tilde{p}(\mathbf{x})) \nabla_{\boldsymbol{\theta}} \log \tilde{p}(\mathbf{x})}{Z}$$

$$= \frac{\sum_{\mathbf{x}} \tilde{p}(\mathbf{x}) \nabla_{\boldsymbol{\theta}} \log \tilde{p}(\mathbf{x})}{Z}$$

$$= \sum_{\mathbf{x}} p(\mathbf{x}) \nabla_{\boldsymbol{\theta}} \log \tilde{p}(\mathbf{x})$$

$$= \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} \nabla_{\boldsymbol{\theta}} \log \tilde{p}(\mathbf{x}).$$

18.1 로그가능도의 기울기

$$\nabla_{\boldsymbol{\theta}} \int \tilde{p}(\mathbf{x}) d\boldsymbol{x} = \int \nabla_{\boldsymbol{\theta}} \tilde{p}(\mathbf{x}) d\boldsymbol{x}.$$

라이프니츠 법칙을 이용해서 다음의 항등식을 얻음

특정한 정칙성 조건들이 충족해야 적용 가능

1. 비정규화 분포가 반드시 θ의 모든 값에 대해 x의 르베그 적분가능(Lebesque-integrable) 함수

2. 기울기 $\nabla_{\boldsymbol{\theta}} \tilde{p}(\mathbf{x})$가 모든 θ와 거의 모든 x에 대해 존재 해야함

3. 모든 θ 와 거의 모든 x에 대해 $\max_i \left| \frac{\partial}{\partial \theta_i} \tilde{p}(\mathbf{x}) \right| \leq R(\boldsymbol{x})$ 라는
의미에서 $\nabla_{\boldsymbol{\theta}} \tilde{p}(\mathbf{x})$ 를 유계로 만드는 적분가능함수R(x)가 존재해야 함

18.1 로그가능도의 기울기

$$\nabla_\theta \log Z = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} \nabla_\theta \log \tilde{p}(\mathbf{x}).$$

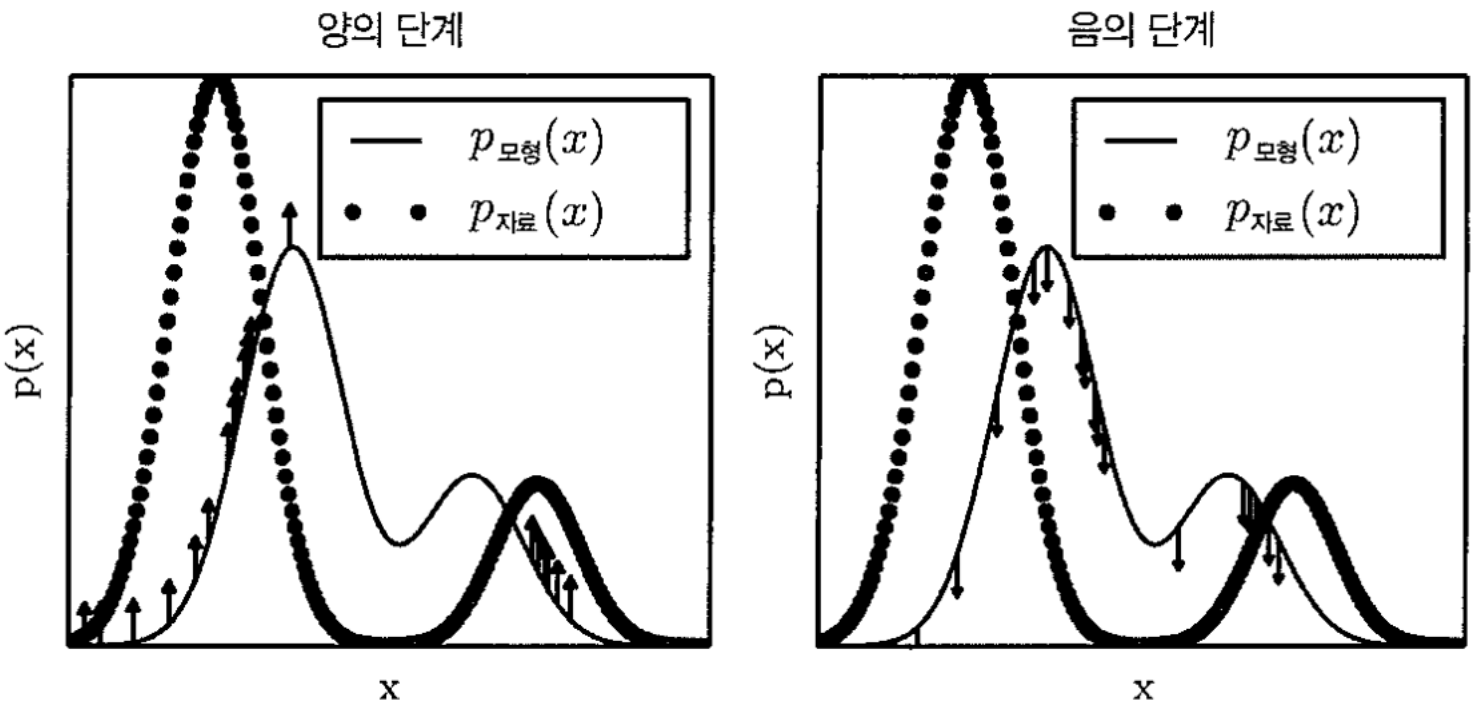
처리 불가능한 분배함수를 가진 모형가의 능도를 근사적으로
최적화하는 **몬테카를로의 한 변형의 기초**

몬테카를로의 접근방식
-> 양의 단계와 음의 단계 모두의 고찰에 도움이 됨

양의 단계 -> 자료에서 뽑은 x에 대한 $\log \tilde{p}(\mathbf{x})$ 증가

음의 단계 -> 자료에서 뽑은 x에 대한 $\log \tilde{p}(\mathbf{x})$ 감소

18.1 로그가능도의 기울기



18.2 확률적 최대가능도와 대조 발산

알고리즘 18.1 처리 불가능한 분배함수를 가진 모형의 로그가능도를 경사 하강법을 이용해서 최대화하는 단순한 MCMC 알고리즘.

단계 크기 ϵ을 작은 양수로 설정한다.

기브스 표집 갱신 단계 수 k를 마르코프 연쇄들이 연소하기에 충분할 정도의 큰 값으로 설정한다. 작은 이미지 패치에 대한 RBM을 훈련하는 데는 100 정도면 될 것이다.

while 아직 수렴하지 않은 동안 **do**

  훈련 집합에서 견본 m개짜리 미니배치 $\{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(m)}\}$을 뽑는다.

  $\mathbf{g} \leftarrow \frac{1}{m}\sum_{i=1}^{m}\nabla_{\boldsymbol{\theta}}\log\tilde{p}(\mathbf{x}^{(i)};\boldsymbol{\theta})$.

  표본 m개짜리 집합 $\{\tilde{\mathbf{x}}^{(1)},\ldots,\tilde{\mathbf{x}}^{(m)}\}$을 무작위로 초기화한다(무작위한 초기치들은 이를테면 균등분포나 정규분포에서 뽑을 수도 있고, 모형의 주변 분포와 부합하는 주변 확률들을 가진 분포에서 뽑을 수도 있을 것이다).

  for $i=1$ **to** k **do**

    for $j=1$ **to** m **do**

      $\tilde{\mathbf{x}}^{(j)} \leftarrow$ 기브스_갱신$(\tilde{\mathbf{x}}^{(j)})$

    end for

  end for

  $\mathbf{g} \leftarrow \mathbf{g} - \frac{1}{m}\sum_{i=1}^{m}\nabla_{\boldsymbol{\theta}}\log\tilde{p}(\tilde{\mathbf{x}}^{(i)};\boldsymbol{\theta})$.

  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \epsilon\mathbf{g}$.

end while

18.2 확률적 최대가능도와 대조 발산

음의단계 계산

-> 음의 단계 항을 근사하는 방법은 여러가지

-> 음의 단계 항의 계산 비용을 낮추긴 함

-> 음의 단계 항 때문에 모형이 잘못된 지점으로 이동하는 부작용

-> 모형의 분포에서 표본을 추출하는 과정이 관여함

-> 모형이 강하게 믿는 점들을 찾는 과정이라고 생각

18.2 확률적 최대가능도와 대조 발산

대조 발산 알고리즘

(contrastive divergence, CD)

-> 각 단계에서 자료분포에서 추출한 표본들로 마르코프 연쇄를 초기화

알고리즘 18.2 경사 하강법을 최적화 절차로 사용하는 대조 발산 알고리즘

단계 크기 ϵ을 작은 양수로 설정한다.
기브스 표집 갱신 단계 수 k를, $p(\mathbf{x};\boldsymbol{\theta})$에서 표본을 추출하는 마르코프 연쇄를 $p_{\text{자료}}$로 초기화했을 때 마르코프 연쇄가 연소하기에 충분할 정도의 큰 값으로 설정한다. 작은 이미지 패치에 대한 RBM을 훈련하는 데는 1에서 20 정도의 값이면 될 것이다.
while 아직 수렴하지 않은 동안 **do**
  훈련 집합에서 견본 m개짜리 미니배치 $\{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(m)}\}$을 뽑는다.
  $\mathbf{g} \leftarrow \frac{1}{m}\sum_{i=1}^{m}\nabla_{\boldsymbol{\theta}}\log\tilde{p}(\mathbf{x}^{(i)};\boldsymbol{\theta})$.
  for $i=1$ **to** m **do**
    $\tilde{\mathbf{x}}^{(i)} \leftarrow \mathbf{x}^{(i)}$.
  end for
  for $i=1$ **to** k **do**
    for $j=1$ **to** m **do**
      $\tilde{\mathbf{x}}^{(j)} \leftarrow$ 기브스_갱신$(\tilde{\mathbf{x}}^{(j)})$.
    end for
  end for
  $\mathbf{g} \leftarrow \mathbf{g} - \frac{1}{m}\sum_{i=1}^{m}\nabla_{\boldsymbol{\theta}}\log\tilde{p}(\tilde{\mathbf{x}}^{(i)};\boldsymbol{\theta})$.
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \epsilon\mathbf{g}$.
end while

18.2 확률적 최대가능도와 대조 발산

-> 자료집합이 이미 주어졌으므로 자료 분포에서 표본을 얻는 비용은 0

-> 초기에는 자료 분포가 모형 분포에 가깝지 않음
-> 음의 단계 부정확

-> 양의 단계는 모형의 자료 확률을 정확히 증가

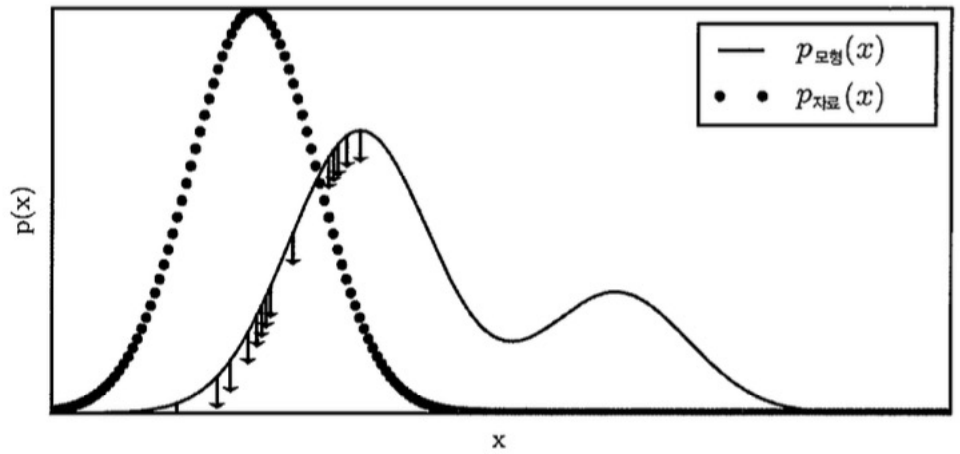
-> 어느 정도 시간이 지나서, 양의 단계가 효과를 발휘하기 시작하면, 모형 분포가 자료 분포에 가까워짐

-> 음의 단계도 정확해짐

18.2 확률적 최대가능도와 대조 발산

대조 발산은 여전히 음의 단계에 대한 근사

가짜 모드(spurious mode)
-> 모형분포에서는 확률이 높지만 자료 생성 분포에서는
확률이 낮은 영역



18.2 확률적 최대가능도와 대조 발산

CD는 RBM 같은 얕은 모형을 훈련하는 데 유용

-> 얕은 모형들을 층층이 쌓아서 DBN, DBM 같은 더 깊은 모형을 초기화 할 수 있음

-> 깊은 모형을 직접 훈련하기는 어려움

-> 가시 단위들의 표본이 주어졌을 때, 은닉 단위들의 표본을 얻기가 어렵기 때문

18.2 확률적 최대가능도와 대조 발산

CD의 문제를 해결하기 위한 방법
-> 경사 하강법의 단계에서 마르코프 연쇄 상태들을 이전 경사 하강법 단계의 것들로 이용해서 초기화

= 확률적 최대가능도(stochastic maximum likelihood , SML)라고 함
(= 지속 대조 발산(persistent contrastive divergence, PCD)

18.2 확률적 최대가능도와 대조 발산

알고리즘 18.3 경사 하강법을 최적화 절차로 사용하는, 확률적 최대가능도와 지속 대조 발산을 조합한 알고리즘.

단계 크기 ϵ을 작은 양수로 설정한다.

기브스 표집 갱신 단계 수 k를, $p(\mathbf{x};\boldsymbol{\theta})$의 표본들로 시작해서 $p(\mathbf{x};\boldsymbol{\theta}+\epsilon\mathbf{g})$에서 표본을 추출하는 마르코프 연쇄가 연소하기에 충분할 정도의 큰 값으로 설정한다. 작은 이미지 패치에 대한 RBM을 훈련하는 데는 1, DBM 같은 좀 더 복잡한 모형에는 5~50 정도가 적당할 것이다.

표본 m개짜리 집합 $\{\tilde{\mathbf{x}}^{(1)},\ldots,\tilde{\mathbf{x}}^{(m)}\}$을 무작위로 초기화한다(무작위한 초기치들은 이를테면 균등분포나 정규분포에서 뽑을 수도 있고, 모형의 주변 분포와 부합하는 주변 확률들을 가진 분포에서 뽑을 수도 있을 것이다).

while 아직 수렴하지 않은 동안 **do**

  훈련 집합에서 견본 m개짜리 미니배치 $\{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(m)}\}$을 뽑는다.

  $\mathbf{g}\leftarrow\frac{1}{m}\sum_{i=1}^{m}\nabla_{\boldsymbol{\theta}}\log\tilde{p}(\mathbf{x}^{(i)};\boldsymbol{\theta})$.

  for $i=1$ **to** k **do**

    for $j=1$ **to** m **do**

      $\tilde{\mathbf{x}}^{(j)}\leftarrow$ 기브스_갱신$(\tilde{\mathbf{x}}^{(j)})$.

    end for

  end for

  $\mathbf{g}\leftarrow\mathbf{g}-\frac{1}{m}\sum_{i=1}^{m}\nabla_{\boldsymbol{\theta}}\log\tilde{p}(\tilde{\mathbf{x}}^{(i)};\boldsymbol{\theta})$.

  $\boldsymbol{\theta}\leftarrow\boldsymbol{\theta}+\epsilon\mathbf{g}$.

end while

18.2 확률적 최대가능도와 대조 발산

-> SML은 CD 보다 가짜모드들을 좀 더 잘 억제함

-> SML 은 표집된 모든 변수 상태를 저장하는 것이 가능함

-> SML 가시단위 뿐만 아니라 은닉단위도 초기화 가능

-> CD는가시단위들만 초기화할 수 있으므로 심층 모형을 위해서는 마르코프 연쇄의 연소가 필요

-> SML은 심층모형을 효율적으로 훈련 할 수 있음

-> RBM에 대해 SML이 가장 좋은 시험 집합 로그가능도를 산출, RBM의 은닉단위들을 SVM 분류기를 위한 특징들로 사용한 경우에 SML이 가장 좋은 결과를 보여줌

18.2 확률적 최대가능도와 대조 발산

SML을 이용해서 훈련한 모형에서 뽑은 표본들을 평가할 때 주의할 점

-> 모형의 훈련을 마친 후 무작위 출발 점으로 초기화한 신선한 마르코프 연쇄로 표본들을 추출해야 한다는 것

-> 훈련에 쓰인 PCD 음의 단계 연쇄에 존재하는 표본들은 이미 모형의 여러 최근 버전들에 영향을 받은 것들이므로, 모형의 수용력이 실제보다 더 크게 나타날 수 있음

18.2 확률적 최대가능도와 대조 발산

MCMC 기반 방법들의 핵심적인 장점

*MCMC는 Markov Chain Monte Carlo의 약어,
몬테 카를로(Monte Carlo) 시뮬레이션과 마르코프 체인(Markov Chain)을 결합한
확률적인 통계적 계산 방법

-> logZ의 기울기 추정값을 제공한다는 것

18. 3 유사가능도

분배함수와 그 기울기를 근사하는 몬테카를로 방법들은
분배함수를 직접 공략

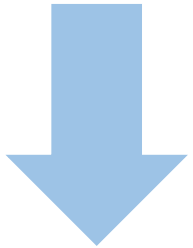
-> 분배함수를 계산하지 않고, 모형을 훈련함으로써 문제를
해소하는 우회적인 방법이 있음

-> 접근방식들은 대부분 무향확률모형에서 확률들의 비(ratio)를
쉽게 계산 할 수 있다는 점에 기초함

-> 확률비를 계산하기 쉬운 이유는 분배함수가 확률 비의 분자와
분모 모두에 나타나기 때문

18. 3 유사가능도

$$\frac{p(\mathbf{x})}{p(\mathbf{y})} = \frac{\frac{1}{Z}\tilde{p}(\mathbf{x})}{\frac{1}{Z}\tilde{p}(\mathbf{y})} = \frac{\tilde{p}(\mathbf{x})}{\tilde{p}(\mathbf{y})}.$$



X가 a와b, c로 분배 된다고 가정

a는 조건부 확률을 알고자 하는 변수들,
b는 그러한 조건부 확률의 조건이 되는 변수들,
c는 조건부확률과 무관한 변수들이라 가정

$$p(\mathbf{a}|\mathbf{b}) = \frac{p(\mathbf{a},\mathbf{b})}{p(\mathbf{b})} = \frac{p(\mathbf{a},\mathbf{b})}{\sum_{\mathbf{a},\mathbf{c}} p(\mathbf{a},\mathbf{b},\mathbf{c})} = \frac{\tilde{p}(\mathbf{a},\mathbf{b})}{\sum_{\mathbf{a},\mathbf{c}} \tilde{p}(\mathbf{a},\mathbf{b},\mathbf{c})}.$$

18. 3 유사가능도

로그가능도를 계산하려면 많은 수의 변수를 주변화 해야함

변수가 n개인 경우 주변화 할 변수는 n-1

확률의 연쇄 법칙에 따라서

$$\log p(\mathbf{x}) = \log p(x_1) + \log p(x_2|x_1) + \ldots + p(x_n|\mathbf{x}_{1:n-1})$$

모든 특징 $\boldsymbol{x}_{-i}$ 이 주어졌을때, 특징 Xi 값을 예측하는 것에 기초한 유사가능도 목적함수

$$\sum_{i=1}^{n} \log p(x_i|\boldsymbol{x}_{-i})$$

$$\sum_{i=1}^{m} \log p(\mathbf{x}_{\mathbb{S}(i)}|\mathbf{x}_{-\mathbb{S}(i)}).$$

일반 유사가능도 함수

18.4 점수 부합과 비 부합

점수 부합

-> z나 그 미분을 추정하지 않고도 모형을 훈련하는 또 다른 일치 추정

인수들에 대한 미분 ∇

$$L(\boldsymbol{x},\boldsymbol{\theta}) = \frac{1}{2}\|\nabla_{\boldsymbol{x}}\log p_{\text{모형}}(\boldsymbol{x};\boldsymbol{\theta}) - \nabla_{\boldsymbol{x}}\log p_{\text{자료}}(\boldsymbol{x})\|_2^2,$$

$$J(\boldsymbol{\theta}) = \frac{1}{2}\mathbb{E}_{p_{\text{자료}}(\boldsymbol{x})}L(\boldsymbol{x},\boldsymbol{\theta}),$$

$$\boldsymbol{\theta}^* = \min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}).$$

-> 입력에 대한 모형의 로그 밀도미분과

입력에 대한 자료의 로그 밀도미분의 기대 제곱 차이를 최소화

18.4 점수 부합과 비 부합

$$L(\boldsymbol{x},\boldsymbol{\theta}) = \frac{1}{2}\|\nabla_{\boldsymbol{x}}\log p_{\text{모형}}(\boldsymbol{x};\boldsymbol{\theta}) - \nabla_{\boldsymbol{x}}\log p_{\text{자료}}(\boldsymbol{x})\|_2^2,$$

$$J(\boldsymbol{\theta}) = \frac{1}{2}\mathrm{E}_{p_{\text{자료}}(\boldsymbol{x})}L(\boldsymbol{x},\boldsymbol{\theta}),$$

$$\boldsymbol{\theta}^* = \min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}).$$

$$\tilde{L}(\boldsymbol{x},\boldsymbol{\theta}) = \sum_{j=1}^{n}\left(\frac{\partial^2}{\partial x_j^2}\log p_{\text{모형}}(\boldsymbol{x};\boldsymbol{\theta}) + \frac{1}{2}\left(\frac{\partial}{\partial x_j}\log p_{\text{모형}}(\boldsymbol{x};\boldsymbol{\theta})\right)^2\right).$$

18.4 점수 부합과 비 부합

비 부합

$$L^{(\mathrm{RM})}(\boldsymbol{x},\boldsymbol{\theta}) = \sum_{j=1}^{n}\left(\frac{1}{1+\frac{p_{\text{모형}}(\boldsymbol{x};\boldsymbol{\theta})}{p_{\text{모형}}(f(\boldsymbol{x}),j);\boldsymbol{\theta})}}\right)^2.$$

->단어 개수 벡터 같은 고차원 희소 자료를 위한 편향으로도 유용

->점수 부합의 기본 착안을 이산자료로 좀 더
성공적으로 일반화한 접근 방식

18.5 잡음 제거 점수 부합

-> 진 분포 p자료가 아닌 다음 분포에 적합시켜서 점수 부합을 정칙화 하는게 바람직한 경우가 있음

$$p_{\text{평활}}(\boldsymbol{x}) = \int p_{\text{자료}}(\boldsymbol{y}) q(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{y}.$$

q(x|y)는 손상과정에 해당

잡음 제거 점수 부합은 실제 응용에서 특히나 유용
-> 실제 응용에서는 진분포p자료에 접근 할 수 없고, 하나의 경험 분포에만 접근할 수 있는 경우가 많음

18.6 잡음 대조 추정

$$\log p_{\text{모형}}(\mathbf{x}) = \log \tilde{p}_{\text{모형}}(\mathbf{x};\boldsymbol{\theta}) + c.$$

$p_{\text{잡음}}(\mathbf{x})$

-> 평가와 표집이 처리 가능한 수준이어야함
-> 이진 분류 변수 y 도입

$$p_{\text{결합}}(y=1) = \frac{1}{2},$$

$$p_{\text{결합}}(\mathbf{x}|y=1) = p_{\text{모형}}(\mathbf{x}),$$

$$p_{\text{결합}}(\mathbf{x}|y=0) = p_{\text{잡음}}(\mathbf{x}).$$

18.6 잡음 대조 추정

$$p_{\text{결합}}(y=1|\mathbf{x}) = \frac{p_{\text{모형}}(\mathbf{x})}{p_{\text{모형}}(\mathbf{x}) + p_{\text{잡음}}(\mathbf{x})}$$

$$= \frac{1}{1 + \frac{p_{\text{잡음}}(\mathbf{x})}{p_{\text{모형}}(\mathbf{x})}}$$

$$= \frac{1}{1 + \exp\left(\log \frac{p_{\text{잡음}}(\mathbf{x})}{p_{\text{모형}}(\mathbf{x})}\right)}$$

$$= \sigma\left(-\log \frac{p_{\text{잡음}}(\mathbf{x})}{p_{\text{모형}}(\mathbf{x})}\right)$$

$$= \sigma(\log p_{\text{모형}}(\mathbf{x}) - \log p_{\text{잡음}}(\mathbf{x})).$$

18.7 분배함수의 추정

-> 무향 그래프 모형과 연관된 처리 불가능한 분배함수Z(θ) 를 계산할 필요가 없는 방법들을 설명

-> 자료의 정규화된 가능도를 계산 해야할 때는 분배함수를 추정할 필요가 있음

$$p_A(\mathbf{x};\boldsymbol{\theta}_A) = \frac{1}{Z_A}\tilde{p}_A(\mathbf{x};\boldsymbol{\theta}_A) \qquad \mathcal{M}_A$$

$$\sum_i \log p_A(\mathrm{x}^{(i)};\boldsymbol{\theta}_A) - \sum_i \log p_B(\mathrm{x}^{(i)};\boldsymbol{\theta}_B) > 0 \qquad \prod_i p_A(\mathrm{x}^{(i)};\boldsymbol{\theta}_A) > \prod_i p_B(\mathrm{x}^{(i)};\boldsymbol{\theta}_B)$$

$\mathcal{M}_A$ 가 더 나은 모형

18.7 분배함수의 추정

$$\sum_i \log p_A(\mathbf{x}^{(i)};\boldsymbol{\theta}_A) - \sum_i \log p_B(\mathbf{x}^{(i)};\boldsymbol{\theta}_B) = \sum_i \left(\log \frac{\tilde{p}_A(\mathbf{x}^{(i)};\boldsymbol{\theta}_A)}{\tilde{p}_B(\mathbf{x}^{(i)};\boldsymbol{\theta}_B)} \right) - m \log \frac{Z(\boldsymbol{\theta}_A)}{Z(\boldsymbol{\theta}_B)}.$$

두 모형의 분배함수들의 비(ratio) 만 알면 되는 형태로 변환해서 상황을 단순하게 만들 수 있음

$$Z_1 = \int \tilde{p}_1(\mathbf{x}) d\mathbf{x}$$

$$= \int \frac{p_0(\mathbf{x})}{p_0(\mathbf{x})} \tilde{p}_1(\mathbf{x}) d\mathbf{x}$$

$$= Z_0 \int p_0(\mathbf{x}) \frac{\tilde{p}_1(\mathbf{x})}{\tilde{p}_0(\mathbf{x})} d\mathbf{x}$$

$$\hat{Z}_1 = \frac{Z_0}{K} \sum_{k=1}^{K} \frac{\tilde{p}_1(\mathbf{x}^{(k)})}{\tilde{p}_0(\mathbf{x}^{(k)})}, \text{ 단 } \mathbf{x}^{(k)} \sim p_0.$$

18.7 분배함수의 추정

중요도 포집 같은 몬테카를로 방법을 이용하면 분배함수를 간단히 추정할 수 있음

$$Z_1 = \int \tilde{p}_1(\mathbf{x}) d\mathbf{x}$$

$$= \int \frac{p_0(\mathbf{x})}{p_0(\mathbf{x})} \tilde{p}_1(\mathbf{x}) d\mathbf{x}$$

$$= Z_0 \int p_0(\mathbf{x}) \frac{\tilde{p}_1(\mathbf{x})}{\tilde{p}_0(\mathbf{x})} d\mathbf{x}$$

$$\hat{Z}_1 = \frac{Z_0}{K} \sum_{k=1}^{K} \frac{\tilde{p}_1(\mathbf{x}^{(k)})}{\tilde{p}_0(\mathbf{x}^{(k)})}, \text{ 단 } \mathbf{x}^{(k)} \sim p_0.$$

비를 추정하는 것도 가능

$$\frac{1}{K} \sum_{k=1}^{K} \frac{\tilde{p}_1(\mathbf{x}^{(k)})}{\tilde{p}_0(\mathbf{x}^{(k)})}, \text{ 단 } \mathbf{x}^{(k)} \sim p_0.$$

18.7 분배함수의 추정

정련된 중요도 표집

-> 정련된 중요도 표집 접근 방식을 이용하면 고차원 공간에 관해 정의된 다봉 분포(훈련된 RBM이 정의하는 분포 같은)의 분배함수를 추정 할 수 있음

$$\frac{Z_1}{Z_0} = \frac{Z_1}{Z_0}\frac{Z_{\eta_1}}{Z_{\eta_1}}\cdots\frac{Z_{\eta_{n-1}}}{Z_{\eta_{n-1}}}$$

$$= \frac{Z_{\eta_1}}{Z_0}\frac{Z_{\eta_2}}{Z_{\eta_1}}\cdots\frac{Z_{\eta_{n-1}}}{Z_{\eta_{n-2}}}\frac{Z_1}{Z_{\eta_{n-1}}}$$

$$= \prod_{j=0}^{n-1}\frac{Z_{\eta_{j+1}}}{Z_{\eta_j}}.$$

18.7 분배함수의 추정

정련된 중요도 표집

-> AIS 전략은 P0에서 표본들을 생성한 후, 전이 연산자들을 이용해서 중간분포들에서 차례로 표본들을 생성하는 과정을 그 표본들이 목표 분포P 의 표본들과 같아질 때까지 반복

- 루프: $k = 1 \ldots K$에 대해
  - $\boldsymbol{x}_{\eta_1}^{(k)} \sim p_0(\mathbf{x})$를 추출한다.
  - $\boldsymbol{x}_{\eta_2}^{(k)} \sim T_{\eta_1}(\mathbf{x}_{\eta_2}^{(k)} | \boldsymbol{x}_{\eta_1}^{(k)})$을 추출한다.
  - ...
  - $\boldsymbol{x}_{\eta_{n-1}}^{(k)} \sim T_{\eta_{n-2}}(\mathbf{x}_{\eta_{n-1}}^{(k)} | \boldsymbol{x}_{\eta_{n-2}}^{(k)})$을 추출한다.
  - $\boldsymbol{x}_{\eta_n}^{(k)} \sim T_{\eta_{n-1}}(\mathbf{x}_{\eta_n}^{(k)} | \boldsymbol{x}_{\eta_{n-1}}^{(k)})$를 추출한다.
- 루프 끝

18.7 분배함수의 추정

다리 표집

-> AIS처럼 중요도 표집의 단점을 해결

$$\frac{Z_1}{Z_0} \approx \sum_{k=1}^{K} \frac{\tilde{p}_*(\boldsymbol{x}_0^{(k)})}{\tilde{p}_0(\boldsymbol{x}_0^{(k)})} \Big/ \sum_{k=1}^{K} \frac{\tilde{p}_*(\boldsymbol{x}_1^{(k)})}{\tilde{p}_1(\boldsymbol{x}_1^{(k)})}.$$

-> p0과 p1이 많이 겹치도록 다리분포 p*를 잘 선택한다면,
다리 표집은 표준적인 중요도 표집보다 거리가 훨씬 더 큰 두 분포의 간극을 메울 수 있음

18.7 분배함수의 추정

다리 표집

연결된 중요도 표집
-> p0과 p1이 너무 멀리 떨어져 있어서 다리 사이의 두 간극을 메울 수 없는 경우라도 AIS로는 다수의 중개분포들을 통해서 두 분포를 연결하는 것이 가능 할 수 있음

18.7 분배함수의 추정

다리 표집

훈련 도중 분배함수 측정
-> 여러 무향 모형의 분배함수추정에서 표준적인 방법으로 자리 잡았지만, 계산비용이 꽤 크기때문에 훈련도중에 사용하기에는 여전히 부적합

-> 이웃연쇄들(병렬 단련에서 온)의 분배함수 비를 다리표집으로 추정한 값들을 시간에 따른 AIS 추정값들과 결합함으로써,
학습과정의 모든 반복에서 분산이 낮은 분배함수 추정값들을 얻을 수 있음