Natural Policy Gradient 와 Fisher Information Matrix

다음 사이트를 참조한다.

- Fisher Information Matrix https://wiseodd.github.io/techblog/2018/03/11/fisher-information/
- Natural Gradient https://wiseodd.github.io/techblog/2018/03/14/natural-gradient/
- keywords
  - A stochastic quasi-newton method

Fisher Information Matrix

Parameter vector θ의 model로 data $x \in \mathbf{R}^n$ 에 대한 확률분포를 $p(x|\theta)$ 라 하자. Maximum likelihood of $p(x|\theta)$를 위한 θ 를 구하는 것이 목표이다. 이때 θ를 Estimation 하기 위해 다음의 Score function을 정의한다.

- Score function : Gradient of log likelihood function

$$s(\theta) = \nabla_\theta \log p(x|\theta) \tag{1}$$

Propostion

The expected value of score function is zero.

Proof. Below, the gradient is w.r.t. θ.

$$
\begin{aligned}
\mathbb{E}_{p(x|\theta)}[s(\theta)] &= \mathbb{E}_{p(x|\theta)}[\nabla \log p(x|\theta)] \\
&= \int \nabla \log p(x|\theta) p(x|\theta) dx \\
&= \int \frac{\nabla p(x|\theta)}{p(x|\theta)} p(x|\theta) dx \\
&= \int \nabla p(x|\theta) dx \\
&= \nabla \int p(x|\theta) dx \\
&= \nabla 1 = 0
\end{aligned}
$$

- 사실, $p(x|\theta)$ 의 거의 대부분의 경우 Gradient는 0이며 Maximum likelihood가 가능한 Local 영역만 Gradient가 0이 되지 않는다. 그러므로 Vanishing gradient는 매우 자연스러운 것이라 볼 수 있다.
- 이 특징을 사용하여 Score 함수의 Covariance를 다음 과 같이 정의한다.

$$
\mathbb{E}_{p(x|\theta)}[(s(\theta) - 0)(s(\theta) - 0)^T] \tag{2}
$$

Fisher Information Matrix

식 (2) 을 **Fishert Information Matrix** 라고 하면 다음과 같이 정의된다.

$$
\mathbf{F} = \mathbb{E}_{p(x|\theta)}[\nabla \log p(x|\theta) \nabla \log p(x|\theta)^T] \tag{3}
$$

Empirical Fisher Information Matrix

실제로 Fisher Information Matrix를 구할 때 Rigorous 하게 구하기는 어려우므로 다음과 같이 단순 평균을 사용하는 경우가 많다. 이 경우를 Empirical Fisher Information이라고 한다.

$$
\mathbf{F} = \frac{1}{N} \sum_{i=1}^{N} \nabla \log p(x_i|\theta) \nabla \log p(x_i|\theta)^T \tag{4}
$$

Fisher and Hessian

- 해당 특징은 완벽히 증명/해석된 것은 아니나, 매우 유력하다.
- Model의 Maximum Likelihood 의 **Expected Hessian** 의 **Negative** 값이다.

Proposition

The negative expected Hessian of log likelihood is equal to the Fisher Information Matrix $\mathbf{F}$.

proof. Log Likelihood 의 Hessian은 Log Likelihood 함수의 Gradient의 Jacobian 과 같으므로

$$
\begin{aligned}
\mathbf{H}_{\log p(x|\theta)} &= \mathbf{J}(\log p(x|\theta)) = \mathbf{J}\left(\frac{\nabla p(x|\theta)}{p(x|\theta)} \right) \\
&= \frac{\mathbf{H}_{p(x|\theta)}}{p(x|\theta)} - \frac{\nabla p(x|\theta) \nabla p(x|\theta)^T}{p(x|\theta) p(x|\theta)} \\
&= \frac{\mathbf{H}_{p(x|\theta)}}{p(x|\theta)} - \left(\frac{\nabla p(x|\theta)}{p(x|\theta)} \right) \left(\frac{\nabla p(x|\theta)}{p(x|\theta)} \right)^T
\end{aligned}
$$

- 위에서 구한 $\mathbf{H}_{\log p(x|\theta)}$의 Expectation value를 구하면

$$
\begin{aligned}
\mathbb{E}_{p(x|\theta)}[\mathbf{H}_{\log p(x|\theta)}] &= \mathbb{E}_{p(x|\theta)}\left[\frac{\mathbf{H}_{\log p(x|\theta)}}{p(x|\theta)} - \left(\frac{\nabla p(x|\theta)}{p(x|\theta)}\right)\left(\frac{\nabla p(x|\theta)}{p(x|\theta)}\right)^T\right] \\
&= \int \frac{\mathbf{H}_{p(x|\theta)}}{p(x|\theta)} p(x|\theta)dx - \mathbb{E}_{p(x|\theta)}[\nabla \log p(x|\theta) \nabla \log p(x|\theta)^T] \\
&= \int \frac{\partial}{\partial x}\frac{\nabla p(x|\theta)}{p(x|\theta)} p(x|\theta)dx - \mathbf{F}, \quad \because \mathbf{F} = \mathbb{E}_{p(x|\theta)}[\nabla \log p(x|\theta) \nabla \log p(x|\theta)^T] \\
&= \frac{\partial}{\partial x}\int \frac{\nabla p(x|\theta)}{p(x|\theta)} p(x|\theta)dx - \mathbf{F} \\
&= \frac{\partial^2}{\partial x^2}\int p(x|\theta)dx - \mathbf{F}, \quad \because \int p(x|\theta)dx = 1 \\
&= -\mathbf{F}
\end{aligned}
$$

- 그러므로 Log Likelihood 함수의 Hessian의 Expectation은 Fisher Information 이다.

$$
\mathbf{F} = -\mathbb{E}_{p(x|\theta)}\mathbf{H}_{\log \mathbf{p}(\mathbf{x}|\theta)} \tag{5}
$$

- 즉, $\mathbf{F}$ as a **measure of curvature of the log likelihood function**.

Summary of Midterm

- Fisher Information Matrix는 **Covariance of the Score function**
- Fisher Information Matrix는**Curvature matrix of a Log Likehood function**.
- Fisher Information Matrix는 **negative expected Hessian of log likelihood function** 이다.
- 따라서 **Optimization에서 Fisher Information Matrix는 Hessian을 대체할 수 있다.**
- Fisher Information Matrix는 **KL Divergence** 와 연관된다.
    - 이는 **Native Gradient**의 유도로 이어진다.

Natural Gradeint Descent

- Likelihood function 자체는 Probability distribution이므로 이러한 공간을 Distribution Space라 칭한다.
    - 따라서, Steepest descent direction in this distribution space를 Parameter space대신으로 생각한다.

KL Divergence

- Reference
    - https://hyunw.kim/blog/2017/10/27/KL_divergence.html
- **Cross Entropy**
    - 확률변수 $\{x_i\}_{i=1}^N$에 대하여 확률분포 $p(x_i), q(x_i)$ 가 존재할 때, Cross Entropy는 다음과 같이 정의된다.

$$
H(p, q) = -\sum_i p(x_i) \log q(x_i) \tag{6}
$$

    - 그런데, Cross Entropy를 전개해 보면 다음과 같이 Entropy가 전개됨을 알 수 있다.

$$
\begin{aligned}
H(p,q) &= -\sum_i p(x_i)\log q(x_i) \\
&= -\sum_i p(x_i)\log q(x_i) - \sum_i p(x_i)\log p(x_i) + \sum_i p(x_i)\log p(x_i) \\
&= -\sum_i (p(x_i)\log q(x_i) - p(x_i)\log p(x_i)) - \sum_i p(x_i)\log p(x_i) \\
&= \sum_i p(x_i)(\log p(x_i) - \log q(x_i)) + H(p), \quad \because H(p) = -\sum_i p(x_i)\log p(x_i) \\
&= \sum_i p(x_i)\log\frac{p(x_i)}{q(x_i)} + H(p)
\end{aligned}
$$

  - 여기에서 KL-Divergence는 다음과 같이 Cross Entropy와 Entropy의 결합으로 설명된다.

$$
KL(p||q) = \sum_i p(x_i)\log\frac{p(x_i)}{q(x_i)} = H(p,q) - H(p) \tag{7}
$$

- 두 Distribution간의 유사성에 대한 Measure
  - Non symmetric 이므로 True Norm or Metric이라 볼 수 없다.
  - 그런데, 일반적인 Loss function $\mathcal{L}$ 에 대하여 다음 식과 같이 정의되므로

$$
\frac{-\nabla_\theta \mathcal{L}(\theta)}{\|\nabla_\theta \mathcal{L}(\theta)\|} = \lim_{\epsilon\to 0}\frac{1}{\epsilon}\underset{d \text{ s.t. } \|d\|\le\epsilon}{\arg\min}\mathcal{L}(\theta + d) \tag{8}
$$

  $d \to 0$ 인 경우 KL-Divergence는 점근적으로 Symmetric이 된다.

- KL Divergence는 만일 전체를 알 수 없는 데이터 set $\{x\}$의 확률분포가 $p(x)$ 로 주어지고, 이 데이터를 모델링하려는 파라미터 θ를 가진 Model의 확률분포를 $q(x|\theta)$ 라하자.
  - 이전의 논의에서는Likelihood $q(x|\theta)$의 (Log) Maximization을 논하였다.
  - 이때, $p(x)$를 따르는x를 N개 추출하는데 다음과 같이 empirical KL Divergence를 놓는다고 하면 (i.e. $p(x_n) = \frac{1}{N}$)

$$
KL(p||q) \approx \frac{1}{N}\sum_{n=1}^{N}[-\log q(x_n|\theta) + \log p(x_n)] \tag{9}
$$

  - 만일, 추출된 x_n이 $p(x_n)$을 따른다고 하면

$$
KL(p||q) \approx \sum_{n=1}^{N} p(x_n)[-\log q(x_n|\theta) + \log p(x_n)] = \sum_{n=1}^{N}[-p(x_n)\log q(x_n|\theta) + p(x_n)\log p(x_n)] = H(p,q) - H(p) \tag{10}
$$

  - 여기에서 $p(x_n)$은 θ에 독립적이므로 변화가 없는 반면
  - $q(x_n|\theta)$ 는 θ를 어떻게 하느냐에 따라 $\log q(x_n|\theta)$를 Maximize 시킬 수 있다.
  - 그런데 $\log q(x_n|\theta)$가 maximize 되면 $KL(p||q)$는 **minimize** 되므로
  - (Log) likelihood mixiimization은 KL Divergence의 Minimization이 된다.

Note

- KL divergence $KL(p||q)$는 p를 기준으로 q와의 Measure가 된다.
  - p는 Data의 분포 q는 데이터 분포로 부터 Model의 분포로 생각할 수 있다.
  - p는 기준점 x_0에서의 분포 q는 $x_0 + d$에서의 분포로 생각할 수 있다.
  - 즉, **Left**에서 **Right**로의 **Divergence** 이다.
- KL Divergence 는 다음과 같이 재 정의된다.

$$
\begin{aligned}
KL(p||q) &= \sum_i p(x_i)\log\frac{p(x_i)}{q_i} \\
&= \sum_i p(x_i)\log p(x_i) - \sum_i p(x_i)\log q_i \\
&= \mathbb{E}_{p(x)}\log p(x) - \mathbb{E}_{p(x)}\log q(x)
\end{aligned}
$$

Fisher Information and KL_Divergence

Proposition : Hessian of KL-Divergence - Fisher Information Matrix

Fisher Information Matrix F is the **Hessian of KL-Divergence** between two distribution $p(x|\theta)$ and $q(x|\theta')$, with respect to θ' , evaluated at $\theta' = \theta$.

proof.

$p(x|\theta)$와 $p(x|\theta')$ 의 KL Divergence의 정의에 의해

$$KL[p(x|\theta)||p(x|\theta')] = \mathbb{E}_{p(x|\theta)}[\log p(x|\theta)] - \mathbb{E}_{p(x|\theta)}[\log p(x|\theta')] \tag{11}$$

θ'에 대한 1차 미분은 ($\theta' = \theta + d$ 이므로 이에 대하여 미분이 되어야 d 에 대한 변화율을 알 수 있다.)

$$\begin{aligned}
\nabla_{\theta'} KL[p(x|\theta)||\, p(x|\theta')] &= \nabla_{\theta'} \mathbb{E}_{p(x|\theta)}[\log p(x|\theta)] - \nabla_{\theta'} \mathbb{E}_{p(x|\theta)}[\log p(x|\theta')] \\
&= -\nabla_{\theta'} \mathbb{E}_{p(x|\theta)}[\log p(x|\theta')] \\
&= -\mathbb{E}_{p(x|\theta)}[\nabla_{\theta'} \log p(x|\theta')] \\
&= -\int p(x|\theta) \nabla_{\theta'} \log p(x|\theta') dx
\end{aligned}$$

θ'에 대한 2차 미분은 따라서 다음과 같이 간단히 구해진다.

$$\begin{aligned}
\frac{\partial}{\partial \theta'} \nabla_{\theta'} KL[p(x|\theta)||p(x|\theta')] &= -\int \left[\frac{\partial}{\partial \theta'} p(x|\theta) \cdot \nabla_{\theta'} \log p(x|\theta') + p(x|\theta) \cdot \frac{\partial}{\partial \theta'} \nabla_{\theta'} \log p(x|\theta') \right] dx \\
&= -\int 0 + p(x|\theta) \frac{\partial^2}{\partial \theta'^2} \log p(x|\theta') dx \\
&= -\int p(x|\theta) \nabla_{\theta'}^2 \log p(x|\theta') dx
\end{aligned}$$

Hessian w.r.t. θ ' evaluated at $\theta' = \theta$ is

$$\begin{aligned}
\mathbf{H}_{KL[p(x|\theta)||p(x|\theta')]} &= -\int p(x|\theta) \nabla_{\theta'}^2 \log p(x|\theta') dx \\
&= -\int p(x|\theta) \mathbf{H}_{\log p(x|\theta')} dx \\
&= -\mathbb{E}_{p(x|\theta)}[\mathbf{H}_{\log p(x|\theta')}] \\
&= \mathbf{F}
\end{aligned}$$

- KL Divergence의 Hessian이 Log Likelihood의 Hessian의 부호만 바뀐 이유는
  - KL Divergence의 한쪽 분포가 데이터에 대한 분포인 관계로 파라미터에 대한 미분에 대해 독립이어서 0으로 떨어지기 때문이다.
  - 따라서, KL Divergencerk 엔트로피 기반인 관계로 (-) 부호가 붙어 KL Divergence의 2차 미분은 Fisher Information Matrixrk 가 된다.
- KL Divergence의 1차 미분은 또한 다음과 같다. (결국, 결론적으로 Maximization of Log Likelihood)

$$\nabla_{\theta'} KL[p(x|\theta)||p(x|\theta')] = -\mathbb{E}_{p(x|\theta)}[\nabla_{\theta'} \log p(x|\theta')] \tag{12}$$

Steepest Descent in Distribution Space

Proposition : Taylor Expansion of KL-Divergence

Let $d \to 0$. The second order Taylor series expansion of KL-Divergence is
$KL[p(x|\theta)||p(x|\theta+d)] \approx \frac{1}{2} d^T F d$

Proof. KL-Divergence의 2차 Taylor 전개는 다음과 같다.

$$\begin{aligned} KL[p_\theta||p_{\theta+d}] &\approx KL[p_\theta||p_\theta] + (\nabla_{\theta'} KL[p_\theta||p_{\theta'}]|_{\theta'=\theta})^T d + \frac{1}{2} d^T \mathbf{H}_{KL[p(x|\theta)||p(x|\theta')]} d \\ &= KL[p_\theta||p_\theta] - \mathbb{E}_{p(x|\theta)}[\nabla_\theta \log p(x|\theta)]^T d + \frac{1}{2} d^T \mathbf{F} d \\ &= \frac{1}{2} d^T \mathbf{F} d, \quad \because KL[p_\theta||p_\theta] = 0,\ \mathbb{E}_{p(x|\theta)}[\nabla_\theta \log p(x|\theta)] = 0 \end{aligned}$$

따라서,

$$KL[p_\theta||p_{\theta+d}] \approx \frac{1}{2} d^T \mathbf{F} d \tag{13}$$

- Loss function $\mathcal{L}(\theta)$ 를 최소화 시키는데 Distribution Space에서 최소화 시키는 것을 생각해보면 다음과 같이 쓸 수 있다.

$$d^* = \underset{d \text{ s.t. } KL[p_\theta||p_{\theta+d}]=c}{\arg\min} \mathcal{L}(\theta+d) \tag{14}$$

  - KL Divergence를 Constant로 고정 시키는 것은 KL Divergence의 변화가 등속도로 이루어지도록 하기 위해서이다.
  - 이를 통해 Curvature의 변화에 무관한 알고리즘을 만들기 위해서이다.
  - 그러나 실제로 최적화 공식을 유도하면 이와 무관한 방정식이 도출된다.
- 식 (13) 을 Largarngian 형식으로 다시 쓰고 여기에 **1차 Taylor Expansion**을 $\mathcal{L}(\theta)$에 적용하면 (KL Divergence에는 2차 Taylor Expansion 적용)

$$\begin{aligned} d^* &= \arg\min_d \mathcal{L}(\theta+d) + \lambda(KL[p_\theta||p_{\theta+d}] - c) \\ &\approx \arg\min_d \mathcal{L}(\theta) + \nabla_\theta \mathcal{L}(\theta)^T d + \lambda \frac{1}{2} d^T \mathbf{F} d - \lambda c \end{aligned}$$

- Minimization 을 구하기 위해 d에 대해 미분하고 미분 값이 0이 되도록 하면

$$\begin{aligned} 0 &= \frac{\partial}{\partial d}\left(\mathcal{L}(\theta) + \nabla_\theta \mathcal{L}(\theta)^T d + \frac{1}{2}\lambda d^T \mathbf{F} d - \lambda c\right) \\ &= \nabla_\theta \mathcal{L}(\theta) + \lambda \mathbf{F} d \\ &\Rightarrow \lambda \mathbf{F} d = -\nabla_\theta \mathcal{L}(\theta) \\ &\Rightarrow d = -\frac{1}{\lambda} \mathbf{F}^{-1} \nabla_\theta \mathcal{L}(\theta) \end{aligned}$$

Definition : Natural Gradient

$$\tilde{\nabla}_\theta \mathcal{L}(\theta) = \mathbf{F}^{-1} \nabla_\theta \mathcal{L}(\theta) \tag{15}$$

Algorithm : Natural Gradient Descent

- Repeat
  - Do forward pass on the model and compute loss $\mathcal{L}$
  - Compute the gradient $\nabla_\theta \mathcal{L}(\theta)$
  - Compute Fisher Information Matrix $\mathbf{F}$ or its emppirical version
  - Compute the natural gradient $\tilde{\nabla}_\theta \mathcal{L}(\theta) = \mathbf{F}^{-1} \nabla_\theta \mathcal{L}(\theta)$
  - Update the parameter: $\theta \leftarrow \theta - \alpha \tilde{\nabla}_\theta \mathcal{L}(\theta)$
- Until Converge

Conclusion

- Big Data에서는 Fisher Information Matrix의 크기가 너무나 커지기 때문에 natural gradient를 사용할 수 없다.
    - 이는 Newton Based 알고리즘이 Deep Learning에서 사용되지 않는 이유와 같다.
- 따라서, Fisher Information Matrix를 Approximation 하는 방법론이 필요하다.
    - ADAM의 경우 그 한 예가 되며 Second moment 가 Fisher Information Matrix를 Diagonal 하게 Approximation 하는 방법이다.

Reference


```
@misc{byrd2015stochastic,
      title={A Stochastic Quasi-Newton Method for Large-Scale Optimization},
      author={R. H. Byrd and S. L. Hansen and J. Nocedal and Y. Singer},
      year={2015},
      eprint={1401.7020},
      archivePrefix={arXiv},
      primaryClass={math.OC}
}
```