Abstract

このノートでは、CD 法の凸性証明、CD 法の勾配の Cross Entropy を用いた表現、および CD 法の非一致性についてまとめました。凸性の照明がメインで、あとはおまけです。

- CD 法の凸性についての証明 (CD 法の評価関数が凸関数であるという論文は探した所、(少なくとも Analytical には) 見つからなかったです。凸性を利用した不等式が幾つか知られているため、それらを使用すれ不偏推定かどうか (あるいはバイアスの性質) が分かるかもしれません。

- CD 法の勾配の式は、任意の指数分布族の場合でも Cross Entropy の形式で書き直す事ができます (CD 法が実際にどのような計算をしているのか理解する助けになるかもしれません)。

- CD 法は一致性 (標本数を増やせば推定値と真値の差はゼロに収束する性質) がない。既に分かっていた事ではありますが記録として残しておきます。

1 CD 法の凸性証明

1.1 準備

この章では凸性の証明の前に、一変数の確率、遷移確率の定義を行います。$p(X, Y|\theta)$ をパラメータ θ に依存する、確率変数 X,Y の同時確率とし, 簡単のため X は Y からの遷移によって得られる確率変数であるとします。

$$p(Y = y) = \frac{1}{|\mathcal{D}|} \sum_{y' \in \mathcal{D}} \delta(y, y') \text{ for } \ y \in \{-1, +1\}^d \tag{1}$$

$$p(X = x|\theta) = \sum_y p(x|y; \theta) p(y|\theta) \tag{2}$$

ここで y が与えられた下での x の条件つき確率 $p(x|y; \theta)$ は y から x への遷移確率に相当し、以下のように定義します。

$$|p(X = x)| = \begin{cases} \Gamma_{xy}(\theta) & y \in \mathcal{D} \text{ and } \|y - x\| = 1 \\ 1 - \sum_{y' | \|y' - x\| = 1} \Gamma_{y'x}(\theta) & y \in \mathcal{D} \text{ and } \|y - x\| = 0 \\ 0 & \text{else} \end{cases} \tag{3}$$

$\Gamma_{xy}(\theta)$ には任意性があります. 例えば以下のように与えるとします。y の状態から Hamming 距離 1 で遷移できる場合の数は d 通りとなります。どの状態が提案されるかは一様確率であるとし

$$\Gamma_{xy}(\theta) := d^{-1} \frac{e^{-\theta g(x)}}{e^{-\theta g(x)} + e^{-\theta g(y)}} = d^{-1}[1 + e^{\theta \Delta g(x,y)}], \ \Delta g(x, y) := g(x) - g(y) \tag{4}$$

すべての確率は状態空間全体で定義されることに注意します（確率分布の定義域が状態空間全体であることから、同時確率の和は交換できます)。

1.2 凸性の証明

前節の定義を踏まえて、CD 法の勾配を書き下して整理します (ただし、$p(Y=y) \to p(y)$ と略記しました)。

$$
\begin{aligned}
\frac{\partial}{\partial\theta}CD(\theta) &= \langle g(x)\rangle^{(0)} - \langle g(1)\rangle^{(1)}_{x|\theta} && (5)\\
&= \sum_y g(y)p^{(0)}(y) - \sum_x g(x)\sum_y p(x|y;\theta)p^{(0)}(y) && (6)\\
&= \sum_y p^{(0)}(y)\{g(y) - \sum_x g(x)p(x|y;\theta)\} && (7)\\
&= \sum_y p^{(0)}(y)\{g(y) - \sum_x g(x)[\ \Gamma_{xy}(\theta)\delta(\|x-y\|-1) + (1-\sum_{y'(\neq x)}\Gamma_{y'x}(\theta))\delta(\|x-y\|)\ ]\} && (8)\\
&= \sum_y p^{(0)}(y)\{g(y) - \sum_x g(x)\delta(\|x-y\|)\} && (9)\\
&\quad -\sum_y p^{(0)}(y)[\ \sum_x g(x)\Gamma_{xy}(\theta)\delta(\|x-y\|-1) - g(y)\sum_{y'(\neq y))}\Gamma_{y'y}(\theta)\ ]\}\\
&= 0 + \sum_y p^{(0)}(y)\sum_x\ [g(y)-g(x)]\ \Gamma_{xy}(\theta)\delta(\|x-y\|-1) && (10)
\end{aligned}
$$

従って、評価関数の勾配は $g(y)-g(x)$、すなわち十分統計量 (sufficient statistics) が遷移確率の重みの役割になっていることがわかります。
$CD(\theta) = \int^\theta$ eq.(10) です。eq(10) の θ 依存性は遷移確率の項のみであるので、知るべきは遷移確率のパラメータ積分です。
eq(4) を用いて、具体的に計算を実行すると以下のようになります。

$$
\begin{aligned}
\int^\theta \Gamma_{xy}(\theta)d\theta' &= \int^\theta d^{-1}\ [1+e^{\theta'(g(x)-g(y))}]\ d\theta' && (11)\\
&= \int^\theta d^{-1}\ [1+e^{\theta'\phi(x,y)}]\ d\theta' && (12)\\
&= -\frac{\log(e^{-\theta\phi(x,y)}+1)}{\phi(x,y)} + \mathrm{const} && (13)
\end{aligned}
$$

eq(10) と eq(13) の結果を合わせると。

$$
\begin{aligned}
CD(\theta) &= \int^\theta d\theta' \sum_y p^{(0)}(y)\sum_x\ [g(y)-g(x)]\ \Gamma_{xy}(\theta')\delta(\|x-y\|-1) && (14)\\
&= \sum_y p^{(0)}(y)\sum_x\ -\phi(x,y)\ \delta(\|x-y\|-1)\int^\theta d\theta'\Gamma_{xy}(\theta') && (15)\\
&= \sum_y p^{(0)}(y)\sum_x\ -\phi(x,y)\ \delta(\|x-y\|-1)\ \{-\frac{\log(e^{-\theta\phi(x,y)}+1)}{\phi(x,y)} + \mathrm{const}\ \} && (16)\\
&= \sum_y\sum_x\ \delta(\|x-y\|-1)\ p^{(0)}(y)\{\log(e^{-\theta\phi(x,y)}+1) + \mathrm{const}\ \} && (17)
\end{aligned}
$$

leq(17) はパラメータ θ に対しての凸関数になります。

(略証)

log が凸関数であること。および、凸関数に非負の係数を掛けて足し合わせた関数もまた凸関数であるからです。$f_i(\xi) \in \{\text{convex function }\}$, に対して、係数を $\alpha_i > 0$, for $i = 1, .., n$ とするとき

$$F(\xi) = \sum_i^n \alpha_i f_i(\xi) \in \{\text{convex function}\} \tag{18}$$

(補足)
関数 f が凸である場合、以下の性質を満たします
$\lambda \in (0, 1), a \leq b$;

$$f(\lambda a + (1-\lambda)b) \leq \lambda f(a) + (1-\lambda) f(b) \tag{19}$$

CD 法の評価関数の凸性は私が調べた限りでは報告がありませんでした (CD 法は Restricted Boltzmann Machine(RB) に使われる例が多いと思うのですが (隠れ変数を使ったボルツマンマシンです)、RB に対しては凸では無いと言っている論文もあったりします)。

1.3 凸性を使用した収束性の証明 (考え中です)

凸性を利用した不等式は数多くあるので、それらの道具を使って不偏推定性 (バイアスの性質) を調べられないかと考えているところです。
いつか見たことがある、不偏性を調べるのに使えそうだなぁと考えているのですが、、この様な不等式をご存知でしたら教えていただきたいです。

$$\|\theta^{(1)} - \theta^{(2)}\| \ \leq \ \|\frac{\partial f(\theta^{(1)})}{\partial \theta} - \frac{\partial f(\theta^{(2)})}{\partial \theta}\| \tag{20}$$

1.4 CD の評価関数とパラメータの事前分布

CD の評価関数は $p(x|y;\theta)$ の形です。今の場合は、パラメータについて natural parameter space で一様確率を仮定している事に相当します。事前分布を仮定して最適化することを認めれば、うまくバイアスが生じない推定ができるのではないかと考えています。(本当は $KL(p^{(0)}\|p^{(\infty)}(\theta) - KL(p^{(1)}\|p^{(\infty)}(\theta)$ と前節で得られた $CD(\theta)$ との差異や他の推定手法との関係を事前分布を介して関連付けしたいのですが、推定手法ごとの評価関数の勾配の違いに確率変数が含まれてきてしまっているため、この考えはダメかもしれません。)

2 CD 法の勾配を Cross Entropy で表現する

勾配を Cross Entropy の表現で書き表します。この書き換えは一般の指数分布族に対して適応できます。

定義:Cross Entropy(交差エントロピー)

$$C(p, q) := -\sum_x p(x) \log q(x) \tag{21}$$

確率分布間の独立性を調べるのに使えそうな entropy 量として Mutual Information(相互情報量) 定義:Mutual Information

$$I(x; y) := \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \tag{22}$$

(当初は、相互情報量で書かれるのではないかと期待していたのですが、そうはならなかったです。)
CD 法評価関数の書き換え

$$
\begin{aligned}
\frac{\partial}{\partial\theta}CD(\theta) &= \langle g(x)\rangle^{(0)} - \langle g(x)\rangle^{(1)}_{x|\theta} && (23)\\
&= \sum_y g(y)p(y) - \sum_y\sum_x g(x)p(x|y;\theta)p(y) && (24)\\
&= \sum_y\sum_x [\ g(y) - g(x)\ ]p(x|y;\theta) && (25)\\
&= \frac{1}{N}\sum_{x^{(n)}\in\mathcal{D}}\sum_{x\in\partial x^{(n)}} [\ g(x^{(n)}) - g(x)\ ]p(x|x^{(n)};\theta) && (26)\\
&= \frac{1}{N}\sum_{x^{(n)}\in\mathcal{D}}\sum_{x\in\partial x^{(n)}} \frac{-1}{\theta'}[\ \log p^{(\infty)}(x^{(n)}|\theta') - \log p^{(\infty)}(x|\theta')\ ]p(x|x^{(n)};\theta) && (27)\\
&= -\frac{1}{N}\frac{1}{\theta'}\sum_{x^{(n)}\in\mathcal{D}}\log p^{(\infty)}(x^{(n)}|\theta') + \frac{1}{\theta'}\sum_x p(x|\theta)\log p^{(\infty)}(x|\theta') && (28)\\
&= \frac{1}{\theta}\{C(p^{(1)}(\ |\theta), p^{(\infty)}) - C(p^{(0)}(\ |\theta), p^{(\infty)})\} && (29)
\end{aligned}
$$

余談ですが、KL-Divergence と Cross Entropy の関係は以下のようになります。

$$
KL(p\|q) = \sum_x p(x)\log\frac{p(x)}{q(x)} = -H(q) + C(p,q) \tag{30}
$$

3 CD 法が一致性を持たないことの証明

最尤推定は、一致性が成り立つにも関わらず、CD 法では一致性が成り立たないことをこの節では書こうと思っているのですが、力尽きてきたためひとまず明日以降にします。
考え方は、CD 法は KL-Divergence の形で書いているため、一致性を示せないんだと思います (以前話亭ことです。)。