Abstract

”Why Contrastive Divergence Works”を読んだ感想（Contrastive Divergence がどのような理屈で、どんな場合に機能するかを議論している論文）についてレポートです (この論文の主張として、Gibbs カーネルを用いた場合などで、状態の更新方法を変えると Maximum Pseudolikelihood Estimation (MPLE), Maximum Composite Likelihood Estimation(MCLE) などに漸近的に一致するため推定できているという主張です (PLME は漸近的に MLE に一致するためバイアスはありません。MCLE のバイアス有無は調べ中です。それぞれの評価関数の真のパラメータ周りの曲率は MLE¿MPLE¿MCLE です))。CD が特にうまくいくときは少ない MC step で同じ変数の更新が繰り返されるときだそうです。
また、CD の moment を導くような評価関数を作る方法を考えました。この評価関数を使用するとパラメータに対しての分散は標本に依らず、0 になります。

1 論文

以下では、大文字の変数を確率変数とします。また、扱う分布は一般的な指数分布であるとします。

$$q(Y) = f(Y)e^{\theta g(Y)+\eta(\theta)}$$

論文中では CD の評価関数として、KL divergence の差を使用するのではなく、別の評価関数を定義し、その評価関数が以下のような評価関数を使用していました。

$$cd(p,q) := \sum_A d_A(p,q)\pi(A) \tag{1}$$

$$\begin{aligned} d_a(p,q) &= KL(p\|q) - KL(p_m\|q_m) + \sum_y \log(\frac{q(Y_a|y_{\setminus a})}{q^*(Y_a|y_{\setminus a}, y^{(0)})})p(y) && (2) \\ &= KL(p\|q * p_m) \geq 0 && (3) \end{aligned}$$

ただし確率変数 y の index を $i \in (1,...,m)$ とし、b を y の index の組みを表すベクトルだとします。まとめた index を $a = \cup_i b_i$ とします。また、$p_m(y_{\setminus a}) = \sum_{y_a} p(y)$ としています。。$d_a(p,q)$ は KL Divergence を用いて、以下のように定義されているとします。

2 簡単な評価関数

以下を CD の評価関数として仮定できると思います。

$$CD(p^{(0)}, p^{(1)}(|\theta^{(s)})\|p^{(\infty)}(\theta)) = KL(p^{(0)}\|p^{(\infty)}(\theta)) - KL(p^{(1)}(|\theta^{(s)})\|p^{(\infty)}(\theta)) \tag{4}$$

$p^{(1)}$ のパラメータ依存性が一つ前の更新時の依る表現が変更点です。CD の評価関数をパラメータ微分すると正しく、moment の式が現れます。また、二階のパラメータ微分をとると、ゼロになります。

$$\frac{\partial}{\partial\theta}CD = \frac{\partial}{\partial\theta}\{\sum_x p^{(0)}(x)\log\frac{p^{(0)}(x)}{p^{(\infty)}(x|\theta)} - \sum_x p^{(1)}(x|\theta^{(s)})\log\frac{p^{(1)}(x|\theta^{(s)})}{p^{(\infty)}(x|\theta)} \tag{5}$$