振り返りと導入

前回は最尤推定量と KL ダイバージェンスを定義した。本稿では次のことを行う:

- KL ダイバージェンスと最尤推定との関係を調べる。
- KL ダイバージェンスの双対平坦多様体への一般化を考える。

1 KL ダイバージェンスと最尤推定

本節では 1 点 x での Dirac 測度を δ^x と記す。

定義 1.1 (Kullback-Leibler ダイバージェンス)**.** 関数 $D\colon \mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{X})\to[0,\infty]$,

$$D(p\|q) := \begin{cases} E_q\left[\frac{dp}{dq}\log\frac{dp}{dq}\right] = E_p\left[\log\frac{dp}{dq}\right] & (p\ll q) \\ \infty & (p\not\ll q) \end{cases} \tag{1.1}$$

を $\mathcal{P}(\mathcal{X})$ 上の **Kullback-Leibler ダイバージェンス** と呼ぶ。

例 1.2 (有限でない例)**.** $p\ll q$ であっても $D(p\|q)<\infty$ とは限らない。$\mathcal{X}:=\mathbb{R}$ として p を標準 Cauchy 分布、q を標準正規分布とした場合が反例のひとつ。

例 1.3 (連続でない例)**.** $\mathcal{P}(\mathcal{X})$ に全変動で定まる位相を入れたとき、KL ダイバージェンスは連続とは限らない。$\mathcal{X}:=\{0,1\}$ として $p_n := \frac{1}{n}\delta^0 + \left(1-\frac{1}{n}\right)\delta^1$, $q_n := \frac{1}{e^n}\delta^0 + \left(1-\frac{1}{e^n}\right)\delta^1$ が反例のひとつ。

$\mathcal{X}=\{1,\dots,n\}, n\in\mathbb{N}$ の場合に最尤推定量と KL ダイバージェンスの関係を考える。

定義 1.4 (経験分布)**.** $x=(x_1,\dots,x_k)\in\mathcal{X}^k$ が与えられたとする。$\mathcal{X}$ 上の確率測度

$$\hat{p}_x := \frac{1}{k}\sum_{j=1}^{k}\delta^{x_j} = \sum_{i=1}^{n} a_i\delta^i, \qquad a_i := \frac{1}{k}\#\{j\in\{1,\dots,k\}\mid x_j=i\} \tag{1.2}$$

を、x により定まる**経験分布 (empirical distribution)** という。

次の定理により、尤度最大化の問題は KL ダイバージェンスの言葉で表せることがわかる。

定理 1.5 (最尤推定量と KL ダイバージェンス)**.** $(\Theta,\mathbf{p})$ を $\mathcal{X}$ 上の統計モデル、$(\Theta,\mathbf{p}^k)$ をその k 個の i.i.d. 拡張とする。$x=(x_1,\dots,x_k)\in\mathcal{X}^k$ が与えられたとし、$\hat{p}_x$ を x により定まる経験分布とする。このとき、集合 $\mathbf{p}(\Theta)$ が $\hat{p}_x$ を支配する確率測度を少なくともひとつ含むならば、次が成り立つ:

$$\operatorname{argmin}_{\theta\in\Theta} D(\hat{p}_x\|\mathbf{p}(\theta)) = \operatorname{argmax}_{\theta\in\Theta} p_\theta^k(x). \tag{1.3}$$

ただし右辺で $p_\theta^k(x) := \prod_{j=1}^{k} p_\theta(x_j)$, $p_\theta(x_j) := \frac{d(\mathbf{p}(\theta))}{di}(x_j)$ (数え上げ測度に関する確率密度関数) である。

証明 定理の仮定より $\exists\theta_0 \in \Theta$ s.t. $\hat{p}_x \ll \mathbf{p}(\theta_0)$ であるが、$(\Theta, \mathbf{p})$ が統計モデルゆえに $\mathbf{p}(\Theta)$ に属する測度はすべて同値だから $\forall\theta \in \Theta$, $\hat{p}_x \ll \mathbf{p}(\theta)$ である。そこで $\forall\theta \in \Theta$ に対し、$\hat{p}_x =: \sum_{i=1}^{n} a_i \delta^i$, $a_i \in [0,1]$ とおくと

$$
\begin{aligned}
D(\hat{p}_x \| \mathbf{p}(\theta)) &= E_{\hat{p}_x}\left[\log \frac{d\hat{p}_x}{d(\mathbf{p}(\theta))}\right] && (1.4)\\
&= \int_{\mathcal{X}} \log \frac{d\hat{p}_x}{d(\mathbf{p}(\theta))}(i)\, d\hat{p}_x(i) && (1.5)\\
&= \sum_{\substack{i\in\mathcal{X}\\ a_i>0}} a_i \log \frac{a_i}{p_\theta(i)} && (1.6)\\
&= -\sum_{\substack{i\in\mathcal{X}\\ a_i>0}} a_i \log p_\theta(i) + C \qquad (C \text{ は } \theta \text{ によらない実定数}) && (1.7)\\
&= -\frac{1}{k}\sum_{j=1}^{k} \log p_\theta(x_j) + C && (1.8)\\
&= -\frac{1}{k} \log p_\theta^k(x) + C && (1.9)
\end{aligned}
$$

ゆえに定理の主張が従う。 □

2 双対平坦多様体の性質

本節では Einstein の記法を用いる。以下、再び一般の $\mathcal{X}$ を考える。

命題 2.1 (指数型分布族と KL ダイバージェンス). $\mathcal{P}$ を $\mathcal{X}$ 上の open な指数型分布族とし、(V, T, μ) を $\mathcal{P}$ の最小次元実現、$\psi\colon \Theta \to \mathbb{R}$ を対数分配関数、$\theta\colon \mathcal{P} \to V^\vee$ を自然パラメータ座標、$\Theta = \theta(\mathcal{P})$ を自然パラメータ空間とする。このとき

$$
D(p\|q) = \psi(\theta(q)) - \psi(\theta(p)) - \frac{\partial\psi}{\partial\theta^i}(p)(\theta^i(q) - \theta^i(p)) \quad (\forall p, q \in \mathcal{P}) \tag{2.1}
$$

が成り立つ。

証明

$$
\begin{aligned}
\psi(\theta(q)) - \psi(\theta(p)) - \frac{\partial\psi}{\partial\theta^i}(p)(\theta^i(q) - \theta^i(p)) &= \psi(\theta(q)) - \psi(\theta(p)) - E_p[T_i](\theta^i(q) - \theta^i(p)) && (2.2)\\
&= E_p\left[\psi(\theta(q)) - \psi(\theta(p)) - \langle\theta(q), T\rangle + \langle\theta(p), T\rangle\right] && (2.3)\\
&= E_p\left[\log \frac{dp}{dq}\right] && (2.4)\\
&= D(p\|q) && (2.5)
\end{aligned}
$$

□

一般の双対平坦多様体にも上の命題の θ, ψ のようなものが存在するかどうかを考える。

定義 2.2 (ポテンシャル). M を多様体、g を M 上の Riemann 計量、∇ を M のアファイン接続とする。$\psi \in C^\infty(M)$ が g の **∇-ポテンシャル** であるとは、$g = \nabla d\psi$ が成り立つことをいう。

定義 2.3 (Hessian チャート)**.** M を多様体、g を M 上の Riemann 計量、∇ を M のアファイン接続とする。M のチャート (U,θ) が ∇ に関する **Hessian チャート** であるとは、θ が ∇-アファイン座標であり、U 上の g の ∇-ポテンシャル $\psi \in C^\infty(U)$ が存在することをいう。

定理 2.4 (Hessian チャートの存在)**.** M を多様体、(g,∇,∇^*) を M 上の双対平坦構造とする。このとき、各 $q \in M$ に対し次が成り立つ:

(1) q のまわりの Hessian チャート (U,θ) が存在する。

(2) U 上の g の ∇-ポテンシャル $\psi \in C^\infty(U)$ をひとつ選んで $\eta_i := \dfrac{\partial \psi}{\partial \theta^i}$ とおくと、$\eta := (\eta_i)_i$ は U 上の ∇^*-アファイン座標であり、(θ,η) は (g,∇,∇^*) に関する双対アファイン座標である。

(3) $\bar\psi := \psi \circ \theta^{-1} \colon \theta(U) \to \mathbb{R}$ の Legendre 変換を $\bar\varphi := \bar\psi^\vee \colon \eta(U) \to \mathbb{R}$ とおき、$\varphi := \bar\varphi \circ \eta \colon U \to \mathbb{R}$ とおくと、φ は U 上の g の ∇^*-ポテンシャルである。

証明 (1) (g,∇,∇^*) が双対平坦ゆえ、とくに q のまわりの ∇-アファイン座標 $\theta = (\theta^i)_i$ が存在する。以下 $\partial_i := \dfrac{\partial}{\partial \theta^i}$ と記す。ここで、g が対称であることと、∇,∇^* が torsion-free であることと、(g,∇,∇^*) が双対構造であることから、$\nabla g \in \Gamma(T^{(0,3)}M)$ は対称である。そこで $h_j := g_{ij}d\theta^i$ $(j=1,\dots,n)$ とおくと、∇g の対称性より h_j は閉形式となるから、Poincaré の補題より局所的に $h_j = d\psi_j$ $(\exists \psi_j)$ と表せる。さらに $h := \psi_j d\theta^j$ とおくと、再び ∇g の対称性より h は閉形式となるから、Poincaré の補題より局所的に $h = d\psi$ $(\exists \psi)$ と表せる。したがって、q の十分小さな開近傍 U が存在し、ψ を ∇-ポテンシャルとして (U,θ) は Hessian チャートとなる。

(2) 以下 $\partial^i := \dfrac{\partial}{\partial \eta_i}$ と記す。まず g の正定値性より $\eta \colon U \to \mathbb{R}^n$ は局所微分同相ゆえ、必要ならば U を小さく取り直して η は微分同相となる。したがって η は U 上の座標となる。また

$$g(\partial_i,\partial^j) = g\left(\partial_i, \frac{\partial \theta^k}{\partial \eta_j}\partial_k\right) = g^{jk}g(\partial_i,\partial_k) = \delta_i^j \tag{2.6}$$

が成り立つから、あとは η が ∇^*-アファイン座標であることを示せばよい。これは次の計算により従う:

$$\begin{aligned}
\Gamma_k^{ij} &= g(\partial_k, \Gamma_l^{ij}\partial^l) && (2.7)\\
&= g(\partial_k, \nabla^*_{\partial^i}\partial^j) && (2.8)\\
&= \partial^i(g(\partial_k,\partial^j)) - g(\nabla_{\partial^i}\partial_k,\partial^j) && (2.9)\\
&= -g(g^{il}\nabla_{\partial_l}\partial_k,\partial^j) && (2.10)\\
&= 0 && (2.11)
\end{aligned}$$

(3) η の定義より $d\psi = \eta_i d\theta^i$ であり、また Legendre 変換の定義より $\psi + \varphi = \theta^i\eta_i$ であることから、$d\varphi = \theta^i d\eta_i$ が成り立つ。したがって

$$\nabla^* d\varphi = \nabla^*(\theta^i d\eta_i) = d\theta^i \otimes d\eta_i = g_{ij}d\theta^i d\theta^j = g \tag{2.12}$$

が成り立つ。よって φ は g の ∇^*-ポテンシャルである。 □

今後の予定

- 双対平坦多様体の canonical ダイバージェンス
- 一般のダイバージェンスと、そこから誘導される双対平坦構造・シンプレクティック構造

- Bayes 更新

参考文献

[Ama16] Shun-ichi Amari, **Information Geometry and Its Applications**, Applied Mathematical Sciences, vol. 194, Springer Japan, Tokyo, 2016 (en).