輪講資料

渡辺 一帆

平成 13 年 10 月 17 日

1　ブラウン運動

最急降下法に確率項を含めたとき，どのような効果が得られるかについて考えよう．確率項を導入するために，まず，実数 $t > 0$ 毎に定義されたユークリッド空間 R^d 上の確率変数たち $\{R_t; t > 0\}$ を考えよう．$D > 0$ を定数とする（拡散係数という）．

ブラウン運動確率変数の集合 $\{R_t; t > 0\}$ が次の 2 条件を満たすとき R^d 上のブラウン運動 (Brownian motion) という．
(1) R_t は R^d 上の確率変数で密度関数

$$p_t(r) = \frac{1}{(4\pi Dt)^{d/2}} \exp(-\frac{\|r\|^2}{4Dt})$$

を持つ（正規分布）．
(2) 任意の自然数 k と任意の $0 < t_1 < t_2 < \cdots < t_k$ について $\{R_{t_{j+1}} - R_{t_j}; j = 1, 2, .., k-1\}$ は独立．

ここで定義した $p_t(r)$ は時刻 0 に原点を出発した「ブラウン運動」が時刻 t に見つかる場所を表現したものである．この密度関数は熱伝導の方程式

$$\frac{\partial p_t(r)}{\partial t} = D\Delta p_t(r)$$

を満たすことが容易に確かめられる．ここで $\Delta = \nabla \cdot \nabla = \sum_{i=1}^{d} (\frac{\partial}{\partial w_i})^2$ はラプラシアンである．

注意確率変数

$$\frac{dR_t}{dt} = \lim_{\tau \to 0} \frac{R_{t+\tau} - R_t}{\tau}$$

は平均が 0 で分散が $(2D/\tau) \to \infty$ の正規分布であって通常の意味では確率変数に収束しないが，$\{dR_t/dt; t > 0\}$ は独立であると考えることができ，ブラウン運動はこの積分と考えられる．$\{dR_t/dt; t > 0\}$ は白色雑音と呼ばれる．

2　確率項を持つ最急降下法

確率項を含む最急降下法を考えよう．$E(w)$ を R^d から R^1 への関数とする．$E(w)$ に関する最急降下法に雑音が入るとどうなるだろうか．単位時間 $\tau > 0$ を定めて，確率変数たち $\{W_k; k = 1, 2, 3, ...\}$ を次のように作る．

$$W_{k+1} = W_k - \tau \nabla E(W_k) + R_\tau \quad (1)$$

$$W_0 = 0 \quad (2)$$

この式は最急降下法に，毎回独立で正規分布に従う雑音 R_τ を加えたものである．

確率項を加えるため W_k は確率変数になるが，どのような性質を持っているだろうか．

$d=1$ の場合について考える．$\partial_w = (\partial/\partial w)$, $\partial_t = (\partial/\partial t)$ の標記を用いる．W_k の密度関数を $q_k(w)$，その特性関数を

$$\varphi_k(s) = \int e^{isw} q_k(w) dw$$

とする．$W_k - \tau \partial_w E(W_k)$ の特性関数は，

$$\int \exp(isw - is\tau\partial_w E(w)) q_k(w) dw$$

であり[1] ，正規分布 R_τ の特性関数は，$\exp(-D\tau s^2)$ である．W_k と R_τ は独立であるから，$W_k - \tau\partial_w E(W_k)$ と R_τ の和の特性関数は，それぞれの積になる．従って，

$$\varphi_{k+1}(s) = \int \exp(isw - is\tau\partial_w E(w) - D\tau s^2) q_k(w) dw$$

この両辺から $\varphi_k(s)$ を減算し，τ が小さな値であるとして，$e^x \cong 1+x$ の近似を用いると，

$$\begin{aligned} \varphi_{k+1}(s) - \varphi_k(s) &= \int [\, \exp(isw)\{\exp(-is\tau\partial_w E(w) - D\tau s^2) - 1\} \,] \, q_k(w) dw \\ &\cong \tau \cdot \int \exp(isw)(-is\partial_w E(w) - Ds^2) q_k(w) dw \end{aligned} \tag{3}$$

フーリエ変換を $\mathcal{F}$ と書くと，一般に $(\mathcal{F}(\partial_w f))(s) = (-is)(\mathcal{F}(f))(s)$ が成り立つので，上式から

$$\begin{aligned} \frac{\varphi_{k+1}(s) - \varphi_k(s)}{\tau} &= (-is)\int \exp(isw)\partial_w E(w) q_k(w) dw + D(-is)^2 \int \exp(isw) q_k(w) dw \\ &= (-is)(\mathcal{F}(\partial_w E(w) q_k(w)))(s) + D(-is)^2 (\mathcal{F} q_k(w))(s) \\ &= (\mathcal{F}[\partial_w(\partial_w E(w) q_k(w))])(s) + D(\mathcal{F}[\partial_w^2 q_k(w)])(s) \end{aligned} \tag{4}$$

ここで逆フーリエ変換を行なうと，

$$\frac{q_{k+1}(w) - q_k(w)}{\tau} = \partial_w(\partial_w E(w) q_k(w)) + D\partial_w^2 q_k(w)$$

が得られる．よって，W_k の密度関数を，時刻 $k\tau$ も含めて $q(w, k\tau)$ と書くと，$q(w, k\tau)$ は偏微分方程式

$$\partial_t q(w,t) - \partial_w(\partial_w E(w) q(w,t)) = D\partial_w^2 q(w,t) \tag{5}$$

を満たす．W_k の次元が $d (d \geq 1)$ の場合に拡張すると，$k\tau = t$, $\tau \to 0$ のとき $q(w,t)$ は偏微分方程式

$$\frac{\partial}{\partial t} q(w,t) - \nabla \cdot (\nabla E(w) q(w,t)) = D\Delta q(w,t) \tag{6}$$

を満たす．この方程式 (6) をフォッカー・プランク方程式 (Fokker-Planck equation) という

時間発展の式 (1) において $\tau \to 0$ の極限操作を（象徴的に）行なって，確率変数に関する微分方程式

$$\frac{dW_t}{dt} = -\nabla E(W_t) + \frac{dR_t}{dt} \tag{7}$$

[1] 確率変数 X の密度関数が $q(x)$ であり，ある確率変数 Y が関数 f によって，$Y = f(X)$ で与えられるとき，Y の特性関数 $\varphi(s)$ は，

$$\varphi(s) = \int_{-\infty}^{\infty} e^{isf(x)} q(x) dx$$

で与えられる．

が得られる．このように確率的な項を含む微分方程式をランジュバン方程式 (Langevin equation) という．この方程式は

$$dW_t = -\nabla E(W_t)dt + dR_t$$

と標記される場合もある．確率微分方程式の差分化である式 (1) は，最急降下法に確率的な項を含めた場合の学習アルゴリズムを表している．計算機によるシミュレーションでは，確率微分方程式自体を行うことは難しいので十分小さな τ を利用する．時間の刻み幅 τ を a 倍すると $\tau\nabla E(w)$ は a 倍になるが，R_τ の「大きさ」は $\sqrt{a}$ 倍になる．

フォッカープランク方程式の解が時刻 $t \to \infty$ で一定の確率密度関数に近づいてゆくとすると，$(\partial q_t/\partial t) = 0$ なので，$\|w\| \to \infty$ で $q_t(w) = 0$ という境界条件を課せば，

$$\begin{aligned}
(6) \quad &\Leftrightarrow \quad -\nabla \cdot (\nabla E(w) q_t(w)) = D\Delta q_t(w) \\
&\Leftrightarrow \quad -(\nabla E(w) q_t(w)) = D\nabla q_t(w) \\
&\Leftrightarrow \quad -\nabla E(w) = D\nabla(\log(q_t(w))) \\
&\Leftrightarrow \quad -E(w) = D\log(q_t(w)) + C \\
&\Leftrightarrow \quad q_t(w) = \frac{1}{C'}\exp(-\frac{1}{D}E(w)) \qquad (8)
\end{aligned}$$

であることがわかる（$C' > 0$ は定数）．

一方，学習モデル $p(x, y|w)$ とパラメータの事前分布 $\varphi(w)$ が与えられたとき，

$$L_n(w) = -\frac{1}{n}\{\sum_{i=1}^{n} \log p(x_i, y_i|w) + \log \varphi(w)\}$$

のように学習誤差 $L_n(w)$ をとると，ベイズ推定を行ったときの事後確率密度関数は

$$\begin{aligned}
p(w|X^n, Y^n) &= \frac{1}{Z}\varphi(w)\prod_{i=1}^{n} p(x_i, y_i|w) \\
&= \frac{1}{Z}\exp(-nL_n(w)) \qquad (9)
\end{aligned}$$

となるが，これは，(8) 式において，$E(w) = nL_n(w)$,$D = 1$ とすれば実現できることになる．