ガウス定常過程に対するスツリングマッチングとデータ圧縮
String Matching and Data Compression for Gaussian Stationary Processes

久保　仁*　　井原俊輔**
Masashi KUBO　　Shunsuke IHARA

Abstract— Wyner and Ziv studied in 1989 the asymptotic properties of recurrence time of stationary ergodic information sources, and applied the results to obtain optimal data-compression schemes. Since then many information theorists have been interested in the study of asymptotic properties of the probability of string matching of a sequence from a source with a data base and the waiting time for the string matching. We consider lossy cases. We derive theorems, for Gaussian stationary processes, concerning the asymptotic behavior of the probability of string matching and the waiting time for the string matching.

Keywords— string matching, data compression, Gaussian process, large deviation

1　はじめに

Wyner-Ziv [4] は, 情報源からの系列がデータベースにある系列と一致するまでの時間の系列の長さを長くしたときの漸近挙動を調べ, データベースを利用する符号化法の最適性を示した. 以来, string matching する確率と string matching するまでの時間の研究が情報理論研究者の関心を集めている.

最近は歪みを許す場合の研究も進み, Yang-Kieffer [5] は, 有限アルファベットの場合で, i.i.d. と Markov 過程の場合について, 入力系列がデータベースに string matching する確率および string matching するまでの時間についての評価を行なった.

我々はガウス定常過程の場合に, string matching の確率および string matching するまでの時間についての評価を行う. なお, i.i.d. 過程に対しては, 任意のアルファベット空間の場合に [5] と同じ結果を我々の方法で示すことができる.

2　問題

$X = \{X_n\}$, $Y = \{Y_n\}$ はそれぞれ情報源, データベースを表す正則 (純非決定的ともいう) なガウス定常過程で, X と Y は独立とする. このとき, 情報源 X からの sample-path $x_1^N = (x_1, \dots, x_N)$ に対し, データベース $Y_1^N = (Y_1, \dots, Y_N)$ に (二乗平均誤差の意味で) 歪み D で string matching する確率

$$P\Big(\frac{1}{N}\sum_{k=1}^{N}|x_k - Y_k|^2 < D\Big)$$

の $N \to \infty$ における漸近挙動を調べる.

また情報源からの系列がデータベースにおいて string matching するまでの時間

$$L_N(Y, X, D) = \inf\Big\{n \geq 0;\ \frac{1}{N}\sum_{k=1}^{N}|X_k - Y_{n+k}|^2 < D\Big\}$$

の $N \to \infty$ における漸近挙動を調べる.

3　準備

必要な用語と記号の説明をしておく.

定常過程は正則なので X, Y はそれぞれスペクトル密度関数 f, g をもつ. このとき k 次の共分散 γ_k は

$$\gamma_k = E[X_n X_{n+k}] = \int_{-\pi}^{\pi} \exp[ik\lambda] f(\lambda)\, d\lambda,$$

で与えられる. なお, 平均は $E[X_n] = E[Y_n] = 0$ とする.

一般に, 空間 Ω 上の二つの確率分布 μ, ν に対し, 相対エントロピー (ダイバージェンス) $D(\mu\|\nu)$ は

$$D(\mu \parallel \nu) = \begin{cases} \displaystyle\int_\Omega \log\Big[\frac{d\mu}{d\nu}(x)\Big]\, d\mu(x), & \mu \prec \nu \\ \infty, & \mu \not\prec \nu \end{cases}$$

で定義される.

定常過程に対し, 相対エントロピー $D_N(X_1^N \| Y_1^N)$, 相互情報量 $I_N(X_1^N, Y_1^N)$ はそれぞれ

$$D_N(X_1^N \parallel Y_1^N) = D(\mu_X^N \parallel \mu_Y^N),$$

$$I_N(X_1^N, Y_1^N) = D(\mu_{XY}^N \parallel \mu_X^N \times \mu_Y^N),$$

で与えられる, ここで μ_X^N は X_1^N の確率分布, μ_{XY}^N は X_1^N と Y_1^N の同時分布を表す.

単位時間当りの相対エントロピー $\overline{D}(X \parallel Y)$ および相互情報量 $\overline{I}(X, Y)$ はそれぞれ

$$\overline{D}(X \parallel Y) = \limsup_{N\to\infty} \frac{1}{N} D_N(X_1^N \parallel Y_1^N)$$

$$\overline{I}(X, Y) = \limsup_{N\to\infty} \frac{1}{N} I_N(X_1^N, Y_1^N)$$

* 名古屋大学大学院理学研究科
Graduate School of Sciences, Nagoya University
Email: kubo@math.nagoya-u.ac.jp
** 名古屋大学情報文化学部
School of Informatics and Sciences, Nagoya University
Email: ihara@math.nagoya-u.ac.jp

とする.

4 結果

我々の目的は, 2節で述べたガウス定常過程に対して, 以下に述べる二つの定理を示すことである.

$D_0 = E[X_n^2] + E[Y_n^2]$ とし, $0 < D < D_0$ に対して,

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} E[|X_n - W_n|^2] \le D$$

を満たす確率過程 $W = \{W_n\}$ の全体を $\boldsymbol{W}_D$ とおく. $R^*(D)$ を次で定める.

$$R^*(D) = \inf_{W\in\boldsymbol{W}_D} \{\overline{I}(W,X) + \overline{D}(W \parallel Y)\}.$$

第一の目的は, string matching の確率について, 次の定理を証明することである.

定理 1 Y のスペクトル密度関数 g は有界と仮定する. このとき μ_X に関して殆どすべての $x = \{x_n\}$ に対し,

$$\lim_{N\to\infty} \frac{1}{N} \log P\Big(\frac{1}{N}\sum_{k=1}^{N} |x_k - Y_k|^2 < D\Big) = -R^*(D)$$

$(0 < D < D_0)$ が成り立つ.

次に, string matching するまでの時間 $L_N(Y,X,D)$ を考える.

Yang-Kieffer [5] は次の混合条件の下でこの問題を論じた. $Y = \{Y_n\}$ に対して $\boldsymbol{F}_m^n$, $\boldsymbol{G}_k$, $\alpha(k)$ を次のように定める.

$$\begin{aligned}
\boldsymbol{F}_m^n &= \sigma(Y_k;\ m \le k \le n)\\
\boldsymbol{G}_k &= \{(G,F);\ \exists(k,l,m)\\
&\qquad \text{s.t. } k \le l \le m-n,\ G \in \boldsymbol{F}_k^l,\ F \in \boldsymbol{F}_m^\infty\}\\
\alpha(k) &= \sup\{|P(G \mid F) - P(G)|;\ (G,F) \in \boldsymbol{G}_k\}
\end{aligned}$$

$\alpha(k)$ を Y の混合係数と呼び, 特に $\sum_{k=1}^{\infty} \alpha(k) < \infty$ のとき, Y は有限和混合係数をもつという.

定理 2 定理1と同じ仮定の下で, Y は有限和混合係数をもつものとする. このとき, 確率 1 で

$$\lim_{N\to\infty} \frac{1}{N} \log L_N(Y,X,D) = R^*(D)$$

$(0 < D < D_0)$ が成り立つ.

上の定理に登場する定数 $R^*(D)$ が情報量と相対エントロピーを使って表せることを, i.i.d. の場合ではあるが, 最初に示したのは古賀, 有本 [3] である.

ガウス定常過程 X の (単位時間当りの) レート・歪み関数 $R(D)$ については, $R(D) = \overline{I}(X,\widetilde{X})$ を満たすガウス定常過程 $\widetilde{X} = \{\widetilde{X}_n\} \in \boldsymbol{W}_D$ が存在することが知られている. 定理1, 2において, 定常過程 Y が $\widetilde{X}$ と同じ分布の場合に限り $R^*(D) = R(D)$ である. このことは, この場合に限れば, データベースを利用して最適な符号化法が構成できることを意味する.

5 定理1の証明

定理1の証明のあたっては, 次に述べる大偏差定理が重要な役割を果たす.

命題 1 (大偏差定理) 確率変数列 $Z = \{Z_n\}$ を考える.

$$\varphi_n(\theta) = \log E[\exp[\theta Z_n]], \qquad \theta \in \mathbf{R}$$

とおく. $\varphi_n(\theta)$ について $n \to \infty$ で極限を持つとき

$$\varphi(\theta) = \lim_{n\to\infty} \frac{1}{n}\varphi_n(\theta n)$$

と定める. $\varphi(\theta)$ が定義され, しかも $\varphi(\theta)$ が C^1 級である θ の集合を $\boldsymbol{D}$ とし, $\boldsymbol{D}' = \{\varphi'(\theta);\ \theta \in \boldsymbol{D}\}$ と記す. 以下, $\boldsymbol{D}^\circ \ne \emptyset$, $\boldsymbol{D}'^\circ \ne \emptyset$ を仮定しておく. $\psi(\theta) = \theta\varphi'(\theta) - \varphi(\theta)$ と定義すると,

$$\varphi^*(t) = \sup_{\theta\in\boldsymbol{D}} \{\theta t - \varphi(\theta)\}$$

で $t = t^*$ を固定したとき, 上限を実現する $\theta = \theta^*$ は $t^* = \varphi'(\theta^*)$ で与えられる. 半平面 $\Pi = \{x \in \mathbf{R};\ \theta^*(x - t^*) > 0\}$ を考え, $A \in \Pi$ を $\forall\delta > 0$ に対し, $A \cap (t^* - \delta, t^* + \delta) \ne \emptyset$ となる開集合とすると,

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} \log P(Z_n \in A) &= \lim_{n\to\infty} \frac{1}{n} \log P(Z_n \in \overline{A})\\
&= -\psi(\theta^*)
\end{aligned}$$

である.

この大偏差定理の証明は [1] などを参照してほしい.

定理1の証明のためには, 次の二つのことを示せばよい.

(1) 任意にとった $x = \{x_n\}$ を固定し, 定常過程 $Z = \{Z_n\}$ を

$$Z_N = \frac{1}{N}\sum_{n=1}^{N} |x_n - Y_n|^2, \qquad N = 1, 2, \ldots$$

で定義する. これに対し上述の大偏差定理に現れる $\varphi_N(N\theta)$, $\varphi(\theta)$, $\varphi'(\theta)$, $\psi(\theta)$ を実際に計算し, この定理を適用する.

(2) $R^*(D)$ の定義における inf において, 下限を実現する $W^* \in \boldsymbol{W}_D$ を実際に構成して,

$$\overline{I}(W^*, X) + \overline{D}(W^* \parallel Y) = \psi(\theta^*)$$

となることを示す. ここで θ^* は $\varphi'(\theta^*) = D$ を満たすもの.

5.1 定理1の証明 (1)

まず, (1) に述べた確率過程 $Z = \{Z_n\}$ に対し $\varphi_N(\theta)$, $\varphi(\theta)$ などを上述のように定める. 以降 $\theta \in \boldsymbol{D}$ とする.

$Y_1^N = (Y_1, \ldots, Y_N)$ の分布を $N(0, \Gamma_N)$ とすると,

$$\begin{aligned}
\varphi_N(N\theta) = &-\frac{1}{2}\log|I_N - 2\theta\Gamma_N|\\
&+\theta\langle(2\theta(\Gamma_N^{-1} - 2\theta I_N)^{-1} + I_N)x, x\rangle
\end{aligned}$$

となる. ここで, $\langle\cdot,\cdot\rangle$ は $\mathbf{R}^N$ での内積である. $\varphi_N(N\theta)/N$ を $N\to\infty$ で極限をとると, 右辺第1項については極限は

$$-\frac{1}{4\pi}\int_{-\pi}^{\pi}\log[1-4\pi\theta g(\lambda)]\,d\lambda$$

である. 第2項の極限の計算がステップ1の中で最も本質的な部分で, 次に挙げる補題を用いることで系列 $x=(x_n)$ に依存しない極限が存在することがわかる.

補題に先だって関係 $\sim$ を以下のように定める. (Gray [2] 参照) N 次正方行列 $A=(a_{jk})$ に対して, ノルム $|A|$ および $\|A\|$ を

$$|A|^2=\Big(\frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{N}|a_{jk}|^2\Big),$$
$$\|A\|^2=\max_{x\in\mathbf{R}^N}\{\langle A^*Ax,x\rangle;\ |x|\le 1\}$$

のように定義する. 正方行列の列 $\{A_N\}, \{B_N\}$ (A_N, B_N は N 次正方行列) が次の1, 2を満たすとき, $A_N\sim B_N$ と記すことにする.

(1) ある $M<\infty$ があって, すべての N について $\|A_N\|$, $\|B_N\|\le M$.
(2) $N\to\infty$ のとき, $\lim_{N\to\infty}|A_N-B_N|=0$ となる.

補題 1 $X=\{X_n\}$ は純非決定的なガウス定常過程で, そのスペクトル密度関数を f とし, $T_N(h)$ をスペクトル密度関数 h の N 次 Toeplitz 行列とする. このとき, $A_N\sim T_N(h)$ ならば,

$$\lim_{N\to\infty}\frac{1}{N}\langle A_NX,X\rangle=\lim_{N\to\infty}\frac{1}{N}E[\langle T_N(h)X,X\rangle]$$
$$=2\pi\int_{-\pi}^{\pi}h(\lambda)f(\lambda)\,d\lambda\qquad\text{a.s.}$$

となる.

補題の証明には $Z_N=\langle T_N(h)X,X\rangle/N$ として大偏差定理を用いる.

この補題を用いると, $N^{-1}\varphi_N(N\theta)$ の右辺第2項の極限は

$$\lim_{N\to\infty}\frac{1}{N}\theta\langle(2\theta(\Gamma_N^{-1}-2\theta I_N)^{-1}+I_N)x,x\rangle$$
$$=\int_{-\pi}^{\pi}\frac{f(\lambda)}{1-4\pi\theta g(\lambda)}\,d\lambda\qquad\text{a.s.}$$

となることが示され, 最終的に

$$\varphi(\theta)=\lim_{N\to\infty}\frac{1}{N}\varphi_N(N\theta)$$
$$=-\frac{1}{4\pi}\int_{-\pi}^{\pi}\log[1-4\pi\theta g(\lambda)]\,d\lambda$$
$$+\int_{-\pi}^{\pi}\frac{f(\lambda)}{1-4\pi\theta g(\lambda)}\,d\lambda$$

が得られる. あとは簡単な計算で

$$\varphi'(\theta)=\int_{-\pi}^{\pi}\frac{g(\lambda)}{1-4\pi\theta g(\lambda)}\,d\lambda+\int_{-\pi}^{\pi}\frac{f(\lambda)g(\lambda)}{(1-4\pi\theta g(\lambda))^2}\,d\lambda$$
$$\psi(\theta)=\frac{1}{4\pi}\int_{-\pi}^{\pi}\log[1-4\pi\theta g(\lambda)]\,d\lambda$$
$$+\frac{1}{4\pi}\int_{-\pi}^{\pi}\frac{\theta g(\lambda)}{1-4\pi\theta g(\lambda)}\,d\lambda$$
$$+4\pi\int_{-\pi}^{\pi}\frac{\theta^2f(\lambda)g(\lambda)}{(1-4\pi\theta g(\lambda))^2}\,d\lambda$$

が得られる. ここで大偏差定理を適用すると,

$$\varphi^*(D)=\sup_{\theta\in\boldsymbol{D}}\{\theta D-\varphi(\theta)\}$$

において上限をとる $\theta=\theta^*$ は $\varphi'(\theta^*)=D$ より得られ, その θ^* について

$$\lim_{N\to\infty}\frac{1}{N}\log P\Big(\frac{1}{N}\sum_{k=1}^{N}|x_k-Y_k|^2<D\Big)=-\psi(\theta^*)$$

となる. これで (1) は終った.

5.2 定理1の証明 (2)

θ^* は (1) で得られたものと同じものとし, $\theta^*<0$ や $\varphi'(\theta^*)=D$ といったことが判っている. ここで

$$h(\lambda)=\frac{g(\lambda)}{1-4\pi\theta^*g(\lambda)}$$

というスペクトル密度関数を考え, またこのスペクトル密度関数に対応する k 次共分散を α_k とおく.

$h(\lambda)$ をスペクトル密度関数に持つガウス定常過程で X と独立なものを $V=\{V_n\}$ とし, $W^*=\{W_n^*\}$ を V, $\{\alpha_k\}$ を用いて,

$$W_n^*=V_n-2\theta^*\sum_{k=-\infty}^{\infty}\alpha_kX_{n-k}$$

で定義する. 以下で次の二つのことを示す.

(i) $\overline{I}(W^*,X)+\overline{D}(W^*\parallel Y)=\psi(\theta^*)$ を示す.
(ii) $R^*(D)=\overline{I}(W^*,X)+\overline{D}(W^*\parallel Y)$ を示す.

f_{W^*} を W^* のスペクトル密度関数とすると, ガウス定常過程について知られている結果から

$$\overline{I}(W^*,X)=\frac{1}{4\pi}\int_{-\pi}^{\pi}\log\Big[\frac{f_{W^*}(\lambda)}{h(\lambda)}\Big]\,d\lambda$$
$$\overline{D}(W^*\parallel Y)=\frac{1}{4\pi}\int_{-\pi}^{\pi}\Big\{\frac{f_{W^*}(\lambda)}{g(\lambda)}$$
$$-\log\Big[\frac{f_{W^*}(\lambda)}{g(\lambda)}\Big]-1\Big\}\,d\lambda$$

が判る.

上式の $f_{W^*}(\lambda)$ と $h(\lambda)$ を $f(\lambda)$, $g(\lambda)$ で表しまとめると,

$$\overline{I}(W^*,X)+\overline{D}(W^*\parallel Y)$$

$$= \frac{1}{4\pi}\int_{-\pi}^{\pi} \frac{\theta^* g(\lambda)}{1-4\pi\theta g(\lambda)}\, d\lambda + 4\pi\int_{-\pi}^{\pi} \frac{\theta^{*2} f(\lambda) g(\lambda)}{(1-4\pi\theta^* g(\lambda))^2}\, d\lambda + \frac{1}{4\pi}\int_{-\pi}^{\pi} \log[1-4\pi\theta^* g(\lambda)]\, d\lambda = \psi(\theta^*)$$

となる. 簡単な計算で

$$E[|W_n^* - X_n|^2] = \int_{-\pi}^{\pi} \frac{g(\lambda)}{1-4\pi\theta^* g(\lambda)}\, d\lambda + \int_{-\pi}^{\pi} \frac{f(\lambda)}{(1-4\pi\theta^* g(\lambda))^2}\, d\lambda = \varphi'(\theta^*) = D$$

が得られ, $W^* \in \boldsymbol{W}_D$ であることが判る. つまり

$$\inf_{W\in\boldsymbol{W}_D}\{\overline{I}(W,X) + \overline{D}(W \parallel Y)\} \le \overline{I}(W^*,X) + \overline{D}(W^* \parallel Y)$$

であるから, あとは逆向きの不等式を示せば良い.

$x = \{x_n\}$ に対し, $\mathbf{R}^N$ 上の確率分布 $\nu_N^*(\cdot|x)$ を

$$\frac{d\nu_N^*(\cdot|x)}{d\mu_Y^N}(y) = \exp\Big[-\varphi_N(N\theta^*)\Big]\exp\Big[\theta^*\sum_{k=1}^N |x_k - y_k|^2\Big]$$

で定める. このとき勝手な $W \in \boldsymbol{W}_D$ に対して,

$$0 \le \iint_{\mathbf{R}^{2N}} \log\Big[\frac{d\mu_{W|X}^N(\cdot|x)}{d\nu_N^*(\cdot|x)}(y)\Big]\, d\mu_{W|X}^N(y|x) d\mu_X^N(x) = \iint_{\mathbf{R}^{2N}} \log\Big[\frac{d\mu_{W|X}^N(\cdot|x)}{d\mu_Y^N}(y)\Big] - \log\Big[\frac{d\nu_N^*(\cdot|x)}{d\mu_Y^N}(y)\Big]\, d\mu_{W|X}^N(y|x)\, d\mu_X^N(x)$$

という不等式が成り立つ. $d\nu_N^*(\cdot|x)$ の定義より, 後の項は

$$\iint_{\mathbf{R}^{2N}} \log\Big[\frac{d\nu_N^*(\cdot|x)}{d\mu_Y^N}(y)\Big]\, d\mu_{W|X}^N(y|x)\, d\mu_X^N(x) = -\varphi_N(N\theta^*) + \theta^* E\Big[\sum_{k=1}^N |X_k - W_k|^2\Big]$$

$\theta^* < 0$ に注意すれば,

$$\liminf_{N\to\infty} \frac{1}{N}\iint_{\mathbf{R}^{2N}} \log\Big[\frac{d\nu_N^*(\cdot|x)}{d\mu_Y^N}(y)\Big]\, d\mu_{W|X}^N(y|x)\, d\mu_X^N(x) \ge -\varphi(\theta^*) + \theta^* D = \theta^*\varphi'(\theta^*) - \varphi(\theta^*) = \psi(\theta^*) = \overline{I}(W^*,X) + \overline{D}(W^* \parallel Y)\}$$

が判る. また

$$I_N(W,X) + D_N(W \parallel Y) = \iint_{\mathbf{R}^{2N}} \log\Big[\frac{d\mu_{W|X}^N(\cdot|x)}{d\mu_Y^N}(y)\Big]\, d\mu_{W|X}^N(y|x) d\mu_X^N(x)$$

なので, 以上二つから

$$\overline{I}(W^*,X) + \overline{D}(W^* \parallel Y) \le \limsup_{N\to\infty} \frac{1}{N}\{I_N(W,Y) + D_N(W \parallel Y)\} \le \overline{I}(W,Y) + \overline{D}(W \parallel Y)$$

を得る.

6 定理2の証明

定理1と以下の定理を用いればよい.

定理 3 (Yang-Kieffer [5]) X は定常でエルゴード的とする. Y は定常で有限和混合係数を持ち, また X と Y は独立とする.

$$\lim_{N\to\infty} \frac{1}{N}\log[P(\rho_N(x_1^N, Y_1^N) < D)] = -R^*, \qquad \text{a.s.}$$

のとき,

$$\lim_{N\to\infty} \frac{1}{N}\log L_N(Y,X,D) = R^* \qquad \text{a.s.}$$

となる. ここで, ρ_N は歪み測度である.

7 最後に

本報告では情報源とデータベースが独立な場合について議論を行なった。この場合、歪みを許す場合の universal に最適な符合化にはなっていない。今後は情報源とデータベースが独立でない場合についても研究する予定である。

参考文献

[1] A. Dembo and O. Zeitouni, Large Deviation Techniques and Applications. Jones and Bartlett Pub., Boston, MA, 1992.

[2] R. M. Gray, Toeplitz and circulant matrices. Stanford Univ., Tech. Report, 6504–1 (1977).

[3] 古賀弘樹, 有本卓, 共通のデータベースをもつ有歪データ圧縮アルゴリズムの漸近特性. 信学技報, **IT93-118** (1994), 73–78.

[4] A. D. Wyner and J. Ziv, Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. IEEE Trans. Inform. Theory, **IT-35** (1989), 1250–1258.

[5] E.-H. Yang and J. C. Kieffer, On the performance of data compression algorithms based upon string matching. preprint.