3C1-3

シェマモデルとSTDP則の結合による記号過程の創発

Emergence of semiosis by connecting Shcema model and STDP learning rule

谷口忠大*1　　椹木哲夫*1
Tadahiro Taniguchi　　Tetsuo Sawaragi

*1京都大学大学院工学研究科
Graduate School of Engineering, Kyoto University

A novel integrative learning architecture, RLSM with a STDP network is described. This architecture models symbol emergence in an autonomous agent engaged in reinforcement learning tasks. The architecture consists of two constitutional learning architectures: a reinforcement learning schema model (RLSM) and a spike timing-dependent plasticity (STDP) network. RLSM is an incremental modular reinforcement learning architecture. It makes an autonomous agent acquire behavioral concepts incrementally. STDP is a learning rule of neuronal plasticity that is found in cerebral cortices and the hippocampus. We found that STDP enables an autonomous robot to associate auditory input with its obtained behavioral concepts and to select reinforcement learning modules more effectively. Auditory signals that are interpreted based on obtained behavioral concepts are revealed to correspond to "signs" in Peirce's semiotic triad. This integrative learning architecture is evaluated in the context of modular learning.

1. はじめに

人工知能研究において，記号（シンボル）をどう扱うべきかという疑問は，何度となく問われてきた．古典的 AI と呼ばれた研究の流れの中では，人間の記号操作能力こそが知能の本質であると考えられたが，その設計原理では実世界で上手く振る舞うロボットの知能を設計することは困難であることが徐々に明らかになってきた．これに対し，NewAI と呼ばれる一連の研究は記号をその設計思想から外す事で実世界で作動しうる機械知能を設計した [4]．しかし，NewAI が対象とした単純な身体動作のみならず，状況に即した適応的行動，意志決定やコミュニケーション等を含んだより高次知能についての知能研究へ踏み出そうとすると，再びシンボルの問題は避けがたいものとなる．Harnad はシンボルを捨てるのではなく，形式的な記号系をロボットが経験する現実世界と適応的に結びつける必要性を説き，その問題を記号接地問題 (SGP: Symbol Grounding Problem) と呼んだ．しかし，この考え方はトップダウンに設計された記号系がそれと独立に与えられたロボットの身体を通して接地しうるという立場に暗に立っていると言える．これに対し，記号創発問題 (SEP: Symbol Emergence Problem) という立場では，より徹底したボトムアップな知能創発の視点から，ロボットの持つ記号系はロボット自身による環境との身体的相互作用から自生的に組織化されるべきであると考える，いわば構成主義的立場をとる [11]．また，このようなボトムアップな記号化能力・記号の運用能力を計算論的に理解すること，またその起源を構成論的に理解することは自律適応的な機械の設計論としてのみならず人間の認知やコミュニケーションを問う認知科学，情報学等の種々の学問においても重要である．

一方，今までの人工知能研究において，記号という存在を捉える時その性質を広く人間にとっての記号の意味も含めて統一的に深く議論する事はあまり出来ていなかった．Peirce の記号論によると，記号とは客観的に存在するものではなく，解釈者が存在して初めて存在しうる記号過程 (semiosis) そのものである [3]．Peirce によると記号とはサイン（sign），対象（object），解釈項（interpretant）の三項関係 (semiotic) により成り立つ．ここで，サインと対象の結びつきは解釈項によって導かれ，その関係は固定的ではない．記号の問題を考える際には記号の学問である記号論にその示唆を求める事が重要となる．

しかし，記号論の研究においても，記号過程が成立するための前段階となる，解釈者内部における記号解釈の為の仮説形成の構成プロセスは明らかでない．我々は本稿を通じて，強化学習シェマモデル [7] と，近年，新皮質や海馬において発見されている STDP 学習則 [1] を結合させることにより，Peirce の記号過程を構成論的に表現しうる最小限のモデルを提案する．これにより，自律ロボットは完全に自己に閉じた学習を通じてサインを適応的に利用できるようになり，状況に即した行動を取れるようになっていく．

連絡先: 谷口忠大，京都大学大学院工学研究科機械理工学専攻，〒606-8501　京都市左京区吉田本町，075-753-5044，tanichu@groove.mbox.media.kyoto-u.ac.jp

2. 強化学習シェマモデル

強化学習シェマモデル（RLSM: Reinforcement Learning Schema Model) は報酬系，環境のダイナミクスといった状況が時不変な系の切り替わりという形で現れる時に，それらを自律的に分節化し，それぞれに対応した強化学習器を配置することで環境適応を行う学習機構である．

2.1 モジュール型学習器と記号創発

身体的相互作用を通した記号創発をモデリングする際にはモジュール型，もしくはそれに類した構造が自己組織化される学習機構がしばしば用いられる．Wolpert らは小脳における複数内部モデルの獲得機構として MOSAIC を提案した [9]. 谷らは RNNPB を提案し，ロボットに様々な時系列情報を獲得させている [6]．岡田らは DBSOM を提案し行動を形成するアトラクタを SOM 上に分散的に記憶させることに成功している [10]．著者らはこれに対して，明示的かつ累増的にモジュールを獲得することの出来るシェマモデル（Schema Model）の枠組みを提案してきた [11]．このシェマモデルを強化学習に拡張する事で強化学習シェマモデルを定式化することができた [7]. 本稿では Q-learning に基づいた RLSM を用いる．

2.2 RLSM の定式化

RLSM において λ 番目のシェマは少なくとも二つの内部関数を持つ．一つは Q-learning における行動価値関数 Q^{λ} であ

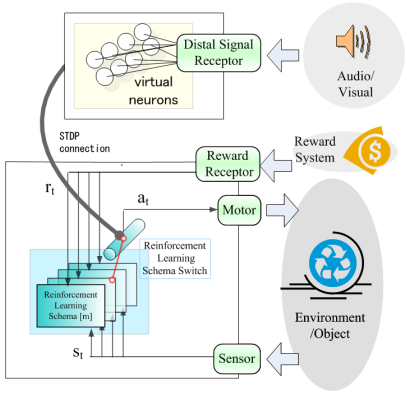

図 1: RLSM with an STDP network

り，もう一つは Q^λ の二次統計量の推定値 $Q^{(2)\lambda}$ である．λ 番目のシェマについての TD 誤差 δ_t，および二次の TD 誤差 $\delta_t^{(2)}$ は以下の式に基づいて算出される．

$$\delta_t^\lambda = r_t + \gamma V^\lambda(s_{t+1}) - Q^\lambda(s_t, a_t) \quad (1)$$

$$\delta_t^{(2)\lambda} = r_t^2 + 2\gamma r_t V^\lambda(s_{t+1}) + \gamma^2 Q^{(2)\lambda}(s_{t+1}, a^*_{t+1}) - Q^{(2)\lambda}(s_t, a_t) \quad (2)$$

$$V^\lambda(s_t) = Q(s_t, a_t^*),\ a_t^* = argmax_a Q^\lambda(s_t, a) \quad (3)$$

$$Q^\lambda(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\delta_t \quad (4)$$

$$Q^{(2)\lambda}(s_t, a_t) \leftarrow Q^{(2)\lambda}(s_t, a_t) + \alpha\delta_t^{(2)} \quad (5)$$

$$\hat{\sigma}^\lambda = \sqrt{Q^{(2)\lambda} - (Q^\lambda)^2}. \quad (6)$$

δ_t を推定された標準偏差 $\hat{\sigma}_t$ により割ることで，無次元量の誤差 R_t^λ を得ることが出来る．これを主観的 TD 誤差と呼ぶ．

$$R^\lambda(t) \equiv |\delta_t^\lambda / \hat{\sigma}_t^\lambda|^2 \quad (7)$$

$$\breve{R}^\lambda(t+\triangle t) = (1-p)R^\lambda(t) + p\breve{R}^\lambda(t) \quad (8)$$

$$\sim \int_0^\infty \frac{1}{\tau} exp(-\frac{s}{\tau}) R^\lambda(t-s) ds \quad (9)$$

$$= \int_{-\infty}^\infty \mathbf{W}_- R^\lambda(t+s) ds \quad (10)$$

$$\mathbf{V}^\lambda(t) = \chi_c(n_p \breve{R}^\lambda(t), n_p),\ n_p = \frac{1+p}{1-p} \quad (11)$$

$$\mathbf{W}_-(t) = \begin{cases} 0 & if\ t > 0 \\ \frac{1}{\tau}\exp(-|t|/\tau) & otherwise \end{cases}$$

ここで $\chi_c(x, n)$ はカイ自乗分布の片側累積関数 $P(X > x)$ である．また，$\breve{R}^\lambda$ は主観的誤差の時間方向についての重み付き平均である．$\breve{*}$ は式 10 によって定義される演算子である．$\mathbf{V}^\lambda$ は λ 番目のシェマの活性度である． シェマ活性度はどれほど現在の環境と報酬系が λ 番目のシェマに適合しているかを表す．p はシェマがどれほど過去の認識を保持するかを表すパラメータである．連続時間では時定数 $\tau = \triangle t/(1-p)$ が p に対応する．ここで $\triangle t$ は連続時間における離散時間の 1 ステップの長さである．さらに，$\breve{R}^\lambda$ は窓関数 $\mathbf{W}_-$ を用いて式 9 のように書き換えられる．有意水準 $\mathbf{V}_\alpha$ が定められると λ 番目のシェマが選択される確率 $P(\lambda)$ は

$$\mu(H^\lambda) = sgn(\mathbf{V}^\lambda(t) - \mathbf{V}_\alpha) \quad (12)$$

$$sgn(x) = \begin{cases} 1 & if\quad x > 0 \\ 0 & otherwise \end{cases} \quad (13)$$

$$P(\lambda) = \mu(H^\lambda) \prod_{k=0}^{\lambda-1} (1 - \mu(H^k)) \quad (14)$$

として定義される．ここで $\mu(H^\lambda)$ は，現在の環境が λ 番目のシェマに反さないという仮説 H^λ の真理値関数である．H^0 は式を簡略化するためのダミー仮説である ($\mu(H^0) = 0$)．もし全てのシェマが選択されなかった場合には新たなシェマが生成され選択される．

2.3 切り替え時間のトレードオフ

RLSM は強化学習シェマを主観的 TD 誤差の重み付け平均 $\breve{R}^\lambda(t)$ に基づいて選択する．しかしながら演算子 $\breve{*}$ が用いる窓関数 $\mathbf{W}_-$ はその重心を $t = -\tau$ に持つ．これはシェマが過去の主観的誤差に基づいてシェマ選択を行っていることを意味する．MOSAIC 等多くのモジュール型学習器にも同様の時間遅れの仕組みが存在する．この時間遅れを解消する為には，τ を小さくすればよいが，小さくしすぎるとモジュール生成・選択が不安定化する．つまり，τ にはシェマ選択の即応性とシェマ生成の安定性についてのトレードオフが存在する．これを解決するためには，状況変化に先行して得られる聴覚情報など別のチャンネルの情報との関係から状況変化を予測する必要がある．次章では STDP 学習則を導入し，STDP がそれを可能にすることを示す．

3. STDP 学習則

近年 STDP(Spike Timing-Dependent Plasticity) 学習則が新皮質や海馬において発見されている [2]． この発見の以前は Hebb 則と呼ばれる学習則が連想学習において主要な役割を果たしていると考えられてきた．しかし，Hebb 則が時間的に対称であり，STDP 則が非対称であるために，Hebb 則で成り立つ理論が STDP 則では多くが成立せず，STDP 則の機能が注目されている． 本章では STDP 則がシェマ選択の遅延解消に寄与し得る事を示す．

3.1 STDP 学習則

神経生物学的実験による STDP 則発見の後に，さまざまな計算論的な STDP のモデルが提案されてきた．一般的には以下のように記述される場合が多い．

$$\triangle w = \sum_{i,o} W(w, t_o^{post} - t_i^{pre}) \quad (15)$$

$$W(w, \triangle t) = \begin{cases} \frac{S_+(w)}{\tau_+}\exp(-|\triangle t|/\tau_+) & if\ t > 0 \\ -\frac{S_-(w)}{\tau_-}\exp(-|\triangle t|/\tau_-) & otherwise \end{cases}$$

ここで $\triangle w$ は pre ニューロンと post ニューロン間の結合の変化量である． t_o^{post} は post ニューロンの o 回目の発火時刻であり，また，t_i^{pre} は pre ニューロンの i 回目の発火時刻である．$W(w, \triangle t)$ は $\triangle t$ について非対称な関数である．τ_+ と τ_- は STDP 学習則の時定数であり，また S_+ と S_- は STDP 学習則の学習ゲインである．一般的にはこれらのゲインパラメータは結合強度 w に依存する．ただし，本稿では結合強度に対してゲインが独立である場合について議論する．また，$\tau_+ = \tau_- = \tau$ とする．

しかし，この学習則のみでは順序だった入出力が与えられると結合強度は発散する．故に w を最大値，最小値の間に制限する hard boundary [5] か，w に比例した減衰項のような付加的なルールを加える soft boundary [1] の設計が必要になる．本稿では w に比例した減衰項を導入する．

$$\triangle w = \sum_i -S_{decay} w(t_i^{pre}). \quad (16)$$

S_{decay} は減衰項のゲインパラメータである．演算子 $\ast$ を用いる事で，全体としての STDP 学習則は単純な力学系として定式化出来る（証明は略す）．

$$\dot{w} = S_+ I\check{O} - S_- O\check{I} - S_{decay} wI, \quad (17)$$

ここで I と O は入力スパイク列と出力スパイク列を表す関数である．

$$I(t) = \sum_{t_i \in \mathcal{T}_I(T)} \delta(t - t_i),\ O(t) = \sum_{t_o \in \mathcal{T}_O(T)} \delta(t - t_o) \quad (18)$$

ここで $\mathcal{T}_I(T)$ と $\mathcal{T}_O(T)$ は pre ニューロンもしくは post ニューロンが発火した時刻の集合である．Dirac のデルタ関数の $\delta(t)$ はスパイクを表わす．

3.2 STDP が結合強度に獲得させるもの

過去についての窓関数 $\mathbf{W}_-$ にたいして，未来についての窓関数 $\mathbf{W}_+$ を以下のように定義する．

$$\mathbf{W}_+(t) = \begin{cases} \frac{1}{\tau}\exp(-|t|/\tau) & if\ t > 0 \\ 0 & otherwise \end{cases}$$

これは窓関数の重心を $t = \tau$ に持つ．次に，未来の平均主観的誤差 $\bar{R}$ を pre ニューロン発火条件下で推定することを試みる．

$$E[\bar{R}|t_i \in \mathcal{T}_I] = E[\int_{-\infty}^{T} \mathbf{W}_+(s) R(t_i + s) ds | t_i \in \mathcal{T}_I] \quad (19)$$

$$\sim E[\int_{-\infty}^{T} (\mathbf{W}_+(s) - \mathbf{W}_-(s)) R(t_i + s) ds | t_i \in \mathcal{T}_I] + \check{R}(t)$$

第一項 T について微分する．

$$\dot{w} = \frac{1}{\#(\mathcal{T}_I(T))}(R\check{I} - \check{R}I - wI) \quad (20)$$

$$= S_+ I\check{R} - S_- R\check{I} - S_{decay} wI \quad (21)$$

ここで $S \equiv S_+ = S_- = S_{decay} = 1/\#(\mathcal{T}_I(T))$ と定義する．これは STDP 学習則が R を post ニューロンの発火と見なす事により未来の主観的誤差平均と過去の主観的誤差平均の差をシナプス結合強度に獲得させられる事を示している．

4. 統合的学習機構による記号創発

STDP 神経回路網の post ニューロンを主観的誤差 R^λ を出力し続ける強化学習シェマで置き換える (図 2)．前章で述べた理由により，STDP 神経回路網は自動的に $\bar{R}^\lambda$ を推定するための情報を i 番目の pre ニューロンと λ 番目のシェマの間の結合強度 w_i^λ に獲得させていく．本稿では，それぞれのニューロンに特有の音が外界で鳴った場合に pre ニューロンが発火するものと考える．RLSM と STDP を組み合わせる事により記号創発を可能にする統合的な学習機構が構築できる．この機構は入ってくる音声入力を自らが自発的に形成したシェマの活性度との関係で理解し，音声を意味づける事を可能とする．窓関数 $\mathbf{W}_+$ はその重心を $t = \tau$ に持つ．故に，RLSM は入ってくる音を “sign” として利用することによって，適切なシェマを時間遅れなく選択することができるようになる．

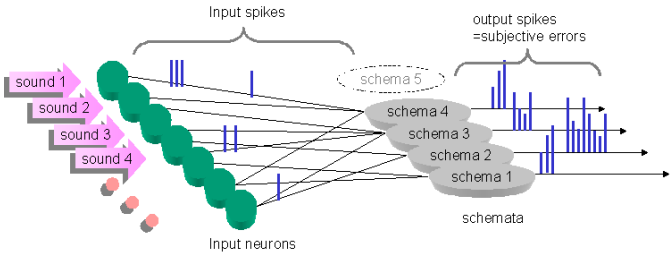

図 2: STDP neuronal network connected to reinforcement learning schemata firing spikes that encode subjective errors

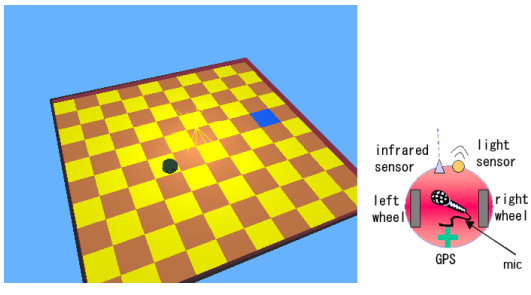

図 3: Left: simulation space; right: Khepera’s sensory-motor system

5. 実験

RLSM に対する STDP 神経回路付加の有効性を移動ロボット Khepera を用いた 2D シミュレータを用いて検証した [8].

5.1 実験環境

図 3 に Khepera のセンサ・モータ系を示す．本実験では cyberbotics 社の Webots を用いる．正方形のシミュレーション空間は長さ 2m の壁に 4 方を囲まれており，光源がその中央 10cm の高さに位置する．Khepera はモータ系として 2 つの車輪を持ち，その回転速度を毎時刻決定することが出来る．また，センサ系として赤外線センサと光センサ，および GPS を持つ．Q-learning が用いる離散状態空間としては Khepera の x，y 座標とその方位角を 6 等分し 216(= 6×6×6) の状態を定義した．また行動空間も 5 つの代表行動に分割した． 赤外線センサの値 ds と光センサの値 ls は 0 と 1 の間の値をとり報酬を計算するためだけに利用する．また，入ってくる音声を得る為にマイクを持つ，それは STDP 神経回路網の pre ニューロンに接続される (図 2)．三つの報酬関数 $r^1 = ds, r^2 = 1.5ls, r^3 = 1.8v_{forward}$ を準備する．ここで $0 \leq v_{forward} \leq 1$ は Khepera の前進速度である．これらの報酬関数を r^1 $(0 < trial \leq 750)$, r^2 $(750 < trial \leq 1500)$, そして r^3 $(1500 < trial \leq 2250)$ の順に切り変える．その後にそれぞれの報酬系は 250 trial ごとに再度切り替えられる．また，STDP についての実験のために 36 種類の音声を仮想的に準備した．3 つのシェマが生成された後に，25 trial 毎に報酬系の切り替えを行った．切り替え時には対応する三つの音声が報酬系 r^1, r^2,r^3 が選ばれた時にそれぞれ鳴らされた．加えて，他の 33 個の音がノイズとしてランダムに鳴らされ続けた． そのようなノイズ環境下で自律ロボットは組織化されたシェマとの関係で意味のある音声を見つけ出さなければならない．もしロボットがその関係性を発見することが出来れば，それらの音声を “sign” として利用することができるようになるかどうかを確かめる．

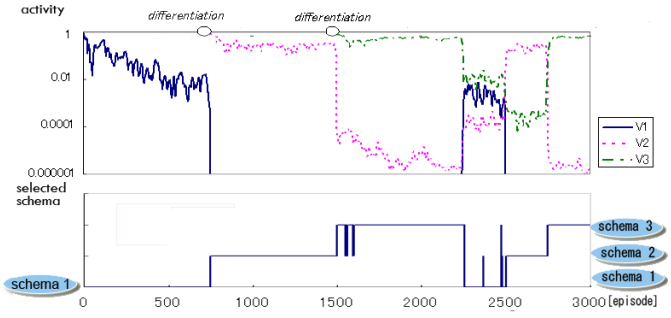

図 4: Top: schema differentiation process and transition of schema activities; bottom: selected reinforcement learning schema

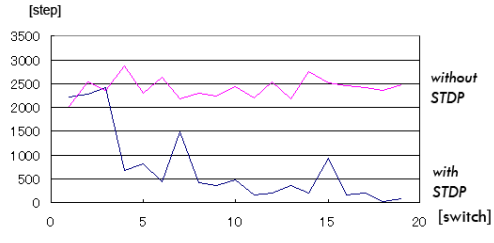

図 5: Course of time delays in switching schemata

5.2 結果

この結果，図 4 のように各シェマ活性度 $\mathbf{V}^{\lambda}$ は遷移した．ここではじめたった一つしか無かった強化学習シェマは三つのシェマに分化し，最終的には 3 つの報酬関数に対応した 3 つの行為概念が組織化された．しかし，この RLSM の検証実験ではシェマの切り替えに平均 2000 step の時間遅れが観察された．これが STDP 学習則による音声のサインとしての利用により解消されていく．

実験の結果，図 6 に示すような結合強度 w が，各シェマと音声の間に得られた．これらの獲得した結合強度を利用することにより Khepera は感知する “sign” をどのシェマを次に選択するかという判断に利用する事が出来るようになった．図 5 に示すように，Khepera が外部報酬系の変化に伴いシェマを切り替えるときに必要とする時間遅れは “sign” を利用することによって徐々に減っていった．この音声というサインの記号としての適応的な利用プロセスが自己閉鎖的な学習から発生するしえる事が示された．これはまさに Peirce の示した記号過程が適応的に可能になっていくプロセスを構成論的に表現したものである..

6. 結言

我々は自律ロボットによる自己閉鎖的な学習プロセスの中で記号創発を実現するための統合的な学習機構を提案した．それは累増的モジュール型強化学習機構 RLSM と時間非対称の自己組織化型学習則 STDP 学習則によって構成される．今後は記号解釈の組織化のみならず，記号を利用してコミュニケーションする自律主体の構成論へと展開していきたい．

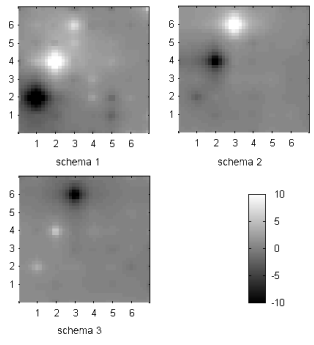

図 6: Obtained synaptic weight on an STDP neuronal network

参考文献

[1] L.F. Abbott and S.B. Nelson. Synaptic plasticity: taming the beast. *nature neuroscience supplement*, Vol. 3, pp. 1178–1182, 2000.

[2] H. Markram, et al. Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps. *SCIENCE*, Vol. 275, pp. 213–215, 1997.

[3] C.S. Peirce, (内田種臣 (編訳)). パース著作集 2　記号学. 勁草書房, 1986.

[4] R. Pfeifer and C. Scheier. *Understanding Intelligence*. The MIT Press (石黒章夫，細田耕，小林宏訳:「知の創成」，共立出版 (2001)), 2001.

[5] S. Song, et al. Competitive hebbian learning through spike-timing-dependent synaptic plasticity. *nature neuroscience*, Vol. 3, pp. 919–926, 2000.

[6] J. Tani, M. Ito, and Y. Sugita. Seif-organization of distributedly represented mulyiple behavior schemata in a mirror system:reviews of robots using rnnpb. *Neural Networks,*, Vol. 17, pp. 1273–1289, 2004.

[7] T. Taniguchi and T. Sawaragi. Incremental acquisition of behavioral concepts through social interactions with a caregiver. In *Artificial Life and Robotics (AROB 11th '06) proceedings*, 2006.

[8] Webots. http://www.cyberbotics.com. Commercial Mobile Robot Simulation Software.

[9] D.M. Wolpert and M. Kawato. Multiple paired forward and inverse models for motor control. *Neural Networks*, Vol. 11, pp. 1317–1329, 1998.

[10] 岡田昌史, 中村大介, 中村仁彦. 力学的情報処理を用いた自己組織的記号獲得と運動生成. 人工知能学会論文誌,, Vol. 20(3), pp. 177–187, 2005.

[11] 谷口忠大, 椹木哲夫. 身体と環境の相互作用を通した記号創発：表象生成の身体依存性についての構成論. システム・制御・情報学会誌, Vol. 18(12), pp. 18–27, 2005.