The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

非定常時系列予測のための確率的勾配法

Concept-drift-aware Stochastic Gradient Descent

宮口航平 *1　梶野洸 *2
Kohei Miyaguchi　Hiroshi Kajino

*1東京大学 The University of Tokyo　*2IBM 東京基礎研究所 IBM Research - Tokyo

We present a stochastic gradient descent for the time-series forecasting problem in the presence of concept drift.

1. 背景と動機

データの生成確率分布が非定常に変化するような，コンセプト・ドリフト環境における時系列予測問題を考える．コンセプト・ドリフト環境においてまず考慮しなければならないことは，環境が変化した際に予測性能が極端に低下しないように，予測をしながらも常に学習を続けるということである．このため，我々は確率的勾配法 (SGD) に基づくオンライン学習の枠組み—データを一つ観測するごとにモデルのパラメタを更新する—に則った方法を考えることにする．SGD に基づくオンライン学習の枠組みにおいて，学習の収束と適応のトレードオフを制御し，予測精度に大きく寄与するパラメタが SGD の学習率である．環境が定常である間は，学習を最適パラメタに向かって収束させるために学習率は小さくなるべきであるが，ひとたび環境が変化したならば，その変化に適応するために学習率を大きく取る必要がある．従って特に，本研究では学習率の自動調整を通じて環境の変化に対応した収束と適応を行う方法について考える．

SGD の学習率自動調整には幾つかの方法が知られているが，我々は特に収束と適応の調整を感度の高いハイパーパラメタなしに行うことができる手法として，variance-based SGD (vSGD) [Schaul 13] に着目する．彼らのアイデアの本質は，最適な学習率が勾配のモーメント統計量を用いて与えられるというものである．このことから，学習率の自動調整問題はそれら統計量の推定問題に帰着される．vSGD の欠点は，各統計量の推定がヒューリスティクスに基づくものであり，推定バイアスが持続してしまうケースが有ることである．本研究ではこの問題を解決するため，より理論的根拠の確かな推定法を導く．

本研究では上記の統計量を推定するアルゴリズムである逐次平均追跡法 (sequential mean tracker, SMT) とともに，コンセプトドリフトのための SGD (concept-drift-aware stochastic gradient descent)，すなわち Cogra を提案する．鍵となるアイデアは，期待予測損失を統計量の推定誤差によって上から抑えることにある．これにより，時系列の予測損失最小化問題が推定量の誤差最小化問題に帰着される．SMT はこの考え方に基づき推定誤差が小さくなる様に各モーメントを推定しており，vSGD のヒューリスティックな推定量を SMT で置き換えることで，Cogra が得られる．

本研究ではさらに，幅広い実験を通して提案法の有効性を検証する．特に次の 3 点を明らかにするため，2 つの実験を行った．**(Q1)** SMT は vSGD の推定器よりもより良くモーメントを推定するか？**(Q2)** SMT の導入は予測性能を改善するか？**(Q3)** どのような場合に Cogra は他の SGD に基づく手法に勝るか？最初の実験では，**(Q1)** に答えるため，人工データを用いてモーメントの推定誤差を計測した．その結果，SMT は vSGD と比較して予測誤差を全体で 60%程度低減することがわかり，**(Q1)** が肯定的に確認された．また 2 つめの実験では，**(Q2)**, **(Q3)** に答えるため，人工データと実データを用いて予測性能を評価した．結果として，Cogra は多くのデータセットにおいて良い予測性能を示し，特に劇的な変化を含むような時系列において他手法を優越する事が分かった．

連絡先: 宮口航平，kohei_miyaguchi@mist.i.u-tokyo.ac.jp

2. 問題設定

$\{X_t\}_{t=1}^{\infty}$ を確率過程とする ($X_t \in \mathcal{X} \subseteq \mathbb{R}^D$)．各時刻 $t = 1, 2, \ldots$ で X_t の実現値 $x_t \in \mathcal{X}$ を観測するとする．各 t に関して，過去のデータを用いて X_t を予測するような確率モデル $p_\theta(X_t \mid x^{t-1})$ を学習したい．ただし，$\theta \in \Theta$ はモデルのパラメタ，そして $x^{t-1} \stackrel{\text{def}}{=} (x_1, x_2, \ldots, x_{t-1})$ は過去の観測値の列を表すとする．データを生成する確率分布は時間とともに変化しうるものであって，ある時刻の最適なパラメタが次の時刻においても最適であるとは限らないので，最適なパラメタの系列 $\theta_1^\star, \theta_2^\star, \ldots$ をオンラインで学習したい：

$$\theta_t^\star = \operatorname*{argmin}_{\theta \in \Theta} \mathcal{L}_t(\theta), \quad \mathcal{L}_t(\theta) \stackrel{\text{def}}{=} \mathbb{E}_{X_t \mid x^{t-1}} \left[\ell(\theta; X_t) \right]. \tag{1}$$

ここで，$\ell(\theta; X_t) \stackrel{\text{def}}{=} -\log p_\theta(X_t \mid x^{t-1})$ は X_t を観測した際にパラメタ θ が受ける損失を表す．以降，損失関数からは X_t を略し，単に $\ell_t(\theta)$ と表記する（添字 t は $X_t \mid x^{t-1}$ の時間変化に依存して損失関数が変化しうることを表す）．

3. 既存手法：vSGD

本節では問題 (1) に対する既存手法である vSGD [Schaul 13] について簡単に説明する．SGD ではパラメタを $\theta_t = \theta_{t-1} - \eta_t \nabla \ell_t(\theta_{t-1})$ と更新するが，vSGD では更新幅 η_t は $\mathcal{L}_t(\theta)$ を貪欲に最小化する様に選ばれる．すなわち，

$$\eta_t^\star = \operatorname*{argmin}_{\eta} \mathbb{E}_{X_t \mid x^{t-1}} \left[\ell_t(\theta_{t-1} - \eta \nabla \ell_t(\theta_{t-1})) \right].$$

これは問題 (1) の実行可能領域 Θ を一次元部分空間 $\{\theta_{t-1} - \eta \nabla \ell_t \in \Theta \mid \eta \in \mathbb{R}\}$ に制限するものであるが，一般には $\eta_t^\star$ は計算することができないので，vSGD では損失関数を二次近似することで $\eta_t^\star = \frac{g_t^2}{h_t v_t}$ という解析解を導

いた．ただし，$g_t \stackrel{\text{def}}{=} \mathbb{E}[\nabla \ell_t(\theta_{t-1})]$，$v_t \stackrel{\text{def}}{=} \mathbb{E}\left[\nabla \ell_t(\theta_{t-1})^2\right]$，$h_t \stackrel{\text{def}}{=} \mathbb{E}[\nabla^2 \ell_t(\theta_{t-1})]$ である．

これらの推定量 g_t, v_t, h_t は一般には未知であるので，過去の勾配の観測から推定を行う．従って，問題 (1) は非定常環境下で勾配のモーメントを推定する問題に帰着される．

具体的には，Schaul らは各推定量を適応的移動平均を用いて推定することを提案した．$\overline{g}_t, \overline{v}_t, \overline{h}_t$ をそれぞれある時刻での $\mathbb{E}[\nabla \ell_t(\theta_{t-1})], \mathbb{E}[(\nabla \ell_t(\theta_{t-1}))^2], \mathbb{E}[(\nabla^2 \ell_t(\theta_{t-1}))^2]$ の推定量とし，τ_t を時定数パラメタとおく．このとき各推定量は次のように更新される:

$$
\begin{aligned}
\overline{g}_t &= (1-\tau_{t-1}^{-1})\overline{g}_t + \tau_{t-1}^{-1}\nabla \ell_t(\theta_{t-1}),\\
\overline{v}_t &= (1-\tau_{t-1}^{-1})\overline{v}_t + \tau_{t-1}^{-1}(\nabla \ell_t(\theta_{t-1}))^2,\\
\overline{h}_t &= (1-\tau_{t-1}^{-1})\overline{h}_t + \tau_{t-1}^{-1}(\nabla^2 \ell_t(\theta_{t-1}))^2,\\
\tau_t &= \left(1-\frac{\overline{g}_t^2}{\overline{v}_t}\right)\tau_{t-1}+1.
\end{aligned}
$$

vSGD は時定数 τ_t を適応的に調整することで，収束と適応のバランスをとる．最適なパラメタが定常である場合（$\theta_t^\star \approx \theta^\star$）には推定されるパラメタは $\theta^\star$ の周辺に集中することになり，$\overline{g}_t$ は 0 に近づく．従って時定数 τ_t は徐々に増加しモーメント推定量の推定が安定化するので，推定されるパラメタの列が収束に向かう．一方，一定の収束期間の後 $\theta_t^\star$ が $\theta^\star$ から変化する時．g_t も同様に 0 から非零へと変化し，遅かれ早かれ $\overline{g}_t$ を非零へと変化させることとなる．これにより時定数が減少に転じることで，vSGD を適応フェーズへと導く．

この τ_t の更新則は一見妥当なものに思えるが．ヒューリスティクスであるがゆえに適応に失敗する場合があり，収束期間が長いほどその後の適応が遅くなる．実際，収束期間が長ければモーメントの比 $\overline{g}_t/\overline{v}_t$ が小さくなり，その間 τ_t が大きくなり続けるため，$\overline{g}_t$ は勾配 $\nabla \ell_t$ の分布の変化を反映しにくくなり，図 1 からも読み取れるように，$\overline{g}_t$ の推定バイアスが長期間に渡って残留し続ける事となる．

4. 提案手法: Cogra

vSGD の枠組みを利用すると，最適学習率問題は勾配のモーメント推定問題に帰着されることがわかった．従って，我々はどのように良い推定量を構成するかということのみに集中すればよい．

良い推定量を構成する上で問題となるのが，背後の分布が変化しうるということである．つまり，どこまでの過去の観測を推定に利用するかを選択しなければならない．分布が変化している時には，最近の観測のみを用いるべきであるし，分布が変化していなければ，より多くの観測データを用いてより精度の高い推定量を計算すべきである．前章で示したように，vSGD は直感的ではあるがヒューリスティックな推定量を用いているため，目的の推定量の推定に失敗することがあり，結果として予測性能をも劣化させることがある．

我々は推定量の良さを新たに定義することでこの問題に対処する．まず各時刻 t で被る予測損失に関して推定量の誤差 $\eta_t^\star$ による上界を導く (4.1 節)．次に，その上界を毎時刻最小化するような移動平均を用いた平均推定オラクルを定義する (4.2 節)．最後に，オラクルを近似する逐次平均追跡法 (SMT) を示し，SMT を用いて最適な学習率を計算することで，非定常環境のための SGD である Cogra を得る．

4.1 リスク：モーメント推定量の良さの指標

$\eta_t^\star$ の代わりに推定量 $\overline{\eta}_t$ を用いてパラメタを更新すると，期待予測損失は定数 C を除いて $\eta_t^\star$ の推定誤差に比例する：

$$
\mathbb{E}_{\overline{\eta}_t}\left[\mathcal{L}_t(\theta_{t-1}-\overline{\eta}_t\nabla \ell_t(\theta_{t-1}))\right] = \frac{hv}{2}r^2(\overline{\eta}_t;\eta_t^\star)+C.
$$

ここで，$r(\overline{\eta}_t;\eta_t^\star) = \sqrt{\mathbb{E}_{\overline{\eta}_t}(\overline{\eta}_t-\eta_t^\star)^2}$ は推定量のリスクを表す．各モーメントの推定量に関してある定数 $M \geq 1$ があって $|\overline{g}_t| \leq M$, $M\overline{v}_t \geq v_t$, $M\overline{h}_t \geq h_t$ が成り立つとする．*1 このとき，$\overline{\eta}_t$ のリスクは個々のモーメント推定量のリスクを用いて抑えることができる：

$$
\begin{aligned}
&r(\overline{\eta}_t;\eta_t^*)\\
&\leq \frac{(g_t+M)M^2}{h_t v_t}r(\overline{g}_t;g_t)+\frac{g_t^2M^2}{h_t v_t^2}r(\overline{v}_t;v_t)+\frac{g_t^2M}{h_t^2 v_t}r(\overline{h}_t;h_t). \quad (2)
\end{aligned}
$$

これは各モーメント推定量のリスクを抑えることが予測誤差の低減につながる事を意味するので，本研究ではリスクをモーメント推定量の良さの指標として採用する．

4.2 貪欲オラクルの存在とリスクの収束

今，誤差を抑えたい統計量はすべてとある確率変数列の各時刻での平均値の形をしているので，確率変数 $Z_t \in \mathbb{R}^d$ の実現値 z_t を逐次的に受け取ってオンラインで平均 $\mu_t = \mathbb{E}[Z_t]$ をリスク $r(\overline{z}_t;\mu_t)$ を小さくするように推定する汎用の平均推定器を考えたい．

具体的には，次のような移動平均を考える：

$$
\overline{z}_0 = \mathbf{0}, \qquad \overline{z}_t = z_t + a_t(\overline{z}_{t-1}-z_t).
$$

ここで $a_t \in \mathbb{R}$ を忘却率と呼ぶ．この方法は 2 つの点で今回の問題設定にふさわしい．一つは，$\overline{z}_t$ が再帰によって定義されるため，計算に有限のメモリしか要しないという点である．もう一つは，忘却率を調節することで過去の観測をどの程度反映するかをコントロールすることができるという点である．

いま，逐次リスク最小化問題

$$
a_t^\star \stackrel{\text{def}}{=} \operatorname*{argmin}_{a\in\mathbb{R}} r^2(\overline{z}_t;\mu_t)
$$

を考える．このとき解はただ一つ存在して，

$$
a_t^\star = \frac{\sigma_t^2-\epsilon_t}{\sigma_t^2+\gamma_{t-1}^2+\delta_t^2-2\epsilon_t} \quad (3)
$$

と書かれる．これをオラクルレートと呼ぶ．ここで，$\sigma_t^2 = \operatorname{tr}\operatorname{Var}(Z_t)$, $\gamma_t^2 = \operatorname{tr}\operatorname{Var}(\overline{z}_t)$, $\delta_t = \|\mathbb{E}[\overline{z}_{t-1}]-\mu_t\|$ and $\epsilon_t = \operatorname{tr}\operatorname{Cov}(\overline{z}_{t-1},Z_t)$ とおいた．

各時刻の局所的な最適性のみに基づいて導かれたオラクルレートはいくつかの望ましい性質を有する．まず，確率変数列 Z_t が互いに独立で平均 μ_t に変化がなければ，$\overline{z}_t$ のリスクは大域的最適なレートで零に収束する．また，Z_t が互いに独立でなく，平均に変化がある場合にでも，次の定理が示される．

定理． $\overline{C}_\tau \stackrel{\text{def}}{=} \sup_{|s-t|\geq\tau}\frac{|\operatorname{tr}\operatorname{Cov}(Z_s,Z_t)|}{\sigma_s\sigma_t}$ とおき，$\overline{C}_1 < 1$ かつ $\overline{C}_\tau \stackrel{\tau\to\infty}{\longrightarrow} 0$ と仮定する． このとき，平均の全変動 $\sum_{t=1}^\infty \|\mu_{t+1}-\mu_t\|$ が有界であって，分散 σ_t^2 が一様に有界である時，オラクルレートを用いた推定量のリスク $r(\overline{z}_t;\mu_t)$ は零に収束する．

*1 これらの仮定は $\overline{g}_t$ に関して適当な上限を設定し，$\overline{v}_t$ と $\overline{h}_t$ を適当な正の下限で下から抑えることで満たされる．

Algorithm 1 Cogra algorithm

Input: Losses $\{\ell_t\}_{t=1}^{\infty}$, initial parameter θ_0, SMT $\mathcal{A}_0$.
Output: Parameters $\{\theta_t\}_{t=1}^{\infty}$.
for each loss function ℓ_t $(t = 1, 2, \ldots)$ **do**
  Incur loss $\ell_t(\theta_{t-1})$
  Receive forgetting rate: $\overline{a}_{t-1} \leftarrow \texttt{ask}(\mathcal{A}_{t-1})$
  Update estimates:
  $\overline{g}_t \leftarrow \nabla\ell_t + \overline{a}_{t-1}(\overline{g}_{t-1} - \nabla\ell_t)$,
  $\overline{v}_t \leftarrow \nabla\ell_t^2 + \overline{a}_{t-1}(\overline{v}_{t-1} - \nabla\ell_t^2)$,
  $\overline{h}_t \leftarrow \nabla^2\ell_t + \overline{a}_{t-1}(\overline{h}_{t-1} - \nabla^2\ell_t)$
  Update SMT: $\mathcal{A}_t \leftarrow \texttt{push}(\mathcal{A}_{t-1}, \nabla\ell_t)$
  Compute learning rate: $\overline{\eta}_t \leftarrow \frac{\overline{g}_t^2}{\overline{h}_t\overline{v}_t}$
  Update parameter: $\theta_t \leftarrow \theta_{t-1} - \overline{\eta}_t\nabla\ell_t(\theta_{t-1})$
end for

この定理は，確率過程 $\{Z_t\}$ がとある混合条件を満たし，推定対象の変化量と分散がそれぞれ有界であるときには，各確率変数が従属していても推定量 $\overline{z}_t$ が真値 μ_t に収束するということをを意味し，ごく一般的な仮定のもとでもオラクルレートを用いた推定量のリスクは零に収束する事がわかる．

4.3 逐次平均追跡法 (SMT)

上で述べたオラクルに基づいて平均 μ_t を推定する方法について示す．オラクルレート $a_t^\star$ は未知の量であるが，式 (3) は確率過程の期待モーメントを用いて書き下すことができることを利用し，期待モーメントを経験的なモーメントで置き換えて近似的にオラクルレートを計算する：

$$\overline{a}_t \stackrel{\text{def}}{=} \frac{\overline{\sigma}_t^2}{\overline{\sigma}_t^2 + \overline{\gamma}_t^2 + \overline{\delta}_t^2}$$

を得る．ここで，ϵ_t については自己相関が十分小さいと仮定して 0 を代入しており，経験モーメントは再帰的に

$$\overline{\sigma}_t^2 = \overline{z^2}_t - \|\overline{z}_t\|^2,\ \overline{\gamma}_t^2 = \overline{\overline{z}^2}_t - \left\|\overline{\overline{z}}_t\right\|^2,\ \overline{\delta}_t = \left\|\overline{\overline{z}}_t - \overline{z}_t\right\|,$$

$$\begin{aligned}\overline{\overline{z}}_t &= \overline{z}_t + \overline{a}_{t-1}(\overline{\overline{z}}_{t-1} - \overline{z}_t),\\ \overline{z^2}_t &= \|z_t\|^2 + \overline{a}_{t-1}(\overline{z^2}_{t-1} - \|z_t\|^2),\\ \overline{\overline{z}^2}_t &= \|\overline{z}_t\|^2 + \overline{a}_{t-1}(\overline{\overline{z}^2}_{t-1} - \|\overline{z}_t\|^2)\end{aligned}$$

と計算される．ただし，最初の忘却率については $\overline{a}_1 = 1/2$ と初期化する．このアルゴリズムを逐次平均追跡法 (SMT) と名付ける．SMT を用いることで平均の推定量の列 $\{\overline{z}_t\}_{t=1}^{\infty}$ をオンラインで計算することができる．

4.4 Cogra

学習率の自動決定に SMT を用いた SGD，Cogra のアルゴリズムを Algorithm 1 に示す．ここで，3 つの推定量 $\overline{g}_t, \overline{v}_t, \overline{h}_t$ に関して，同じ忘却率 $\overline{a}_t$ を用いた．これは同じ関数 $\hat{\mathcal{L}}_t(\theta) = \sum_{s\le t}\prod_{s<i\le t}(1-a_i)a_s\ell_s(\theta)$ の勾配や曲率のモーメントの推定量を用いて最適な学習率を計算することに対応し，それぞれ個別の忘却率を用いる場合と比較して推定を安定させることができると考えられるためである．また，$\overline{g}_t$ の忘却率に統一したのは，式 (2) からわかるように，3 つのリスクのうち収束期において最も支配的な項を減少させるためである．

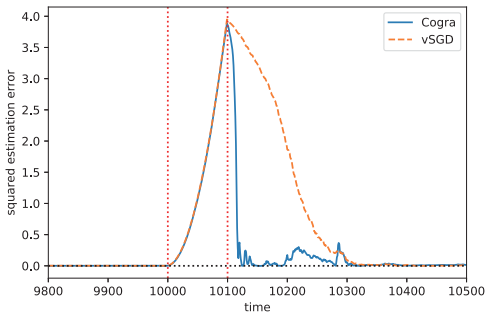


図 1: g_t の推定誤差．x 軸と y 軸はそれぞれ時間と g_t の二乗推定誤差を表す．垂直な点線は変化の開始時刻と終了時刻を表す．Cogra (青実線) は vSGD (橙破線) と比較して有意に早い段階で推定誤差を収束させていることがわかる．

5. 実験

Cogra の有効性について，(i)SMT のモーメント予測の精度および (ii)SMT を採用したことによる予測性能の向上について検証した．すべての実験でベクトル自己回帰モデル (VAR) [Lütkepohl 05] を用いて予測を行った．

5.1 SMT の性能

SMT のモーメントの推定精度を測るため．$d = 1$, $p = 0$, $\mu_t = -1 \to 1$ と徐々に変化させて SMT の推定誤差を vSGD のそれと比較した（図 1）．SMT の方が全体で 60%ほど累積推定誤差が小さい事がわかる．従って，**(Q1)** は肯定的に解決された．

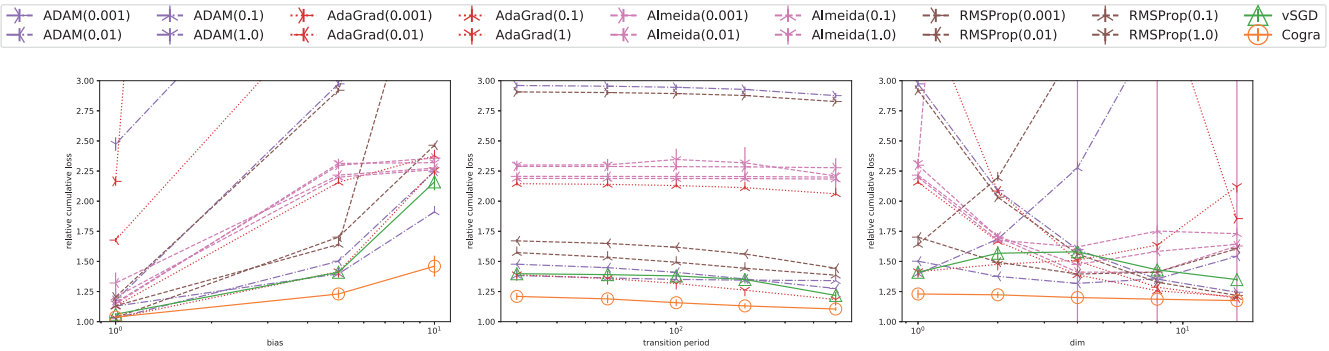
5.2 予測性能

次に，SMT が予測性能にどれほど貢献するかについて人工データと実データを用いて調べた．比較手法は AdaGrad[Duchi 11], ADAM[Kingma 14], Almeida[Almeida 99], RMSProp[Tieleman 12], そして vSGD を採用し，ハイパーパラメタがあるものについては適当なグリッド上でチューニングした．vSGD に関しては τ_t が小さい時に安定しないという既知の問題を避けるため，$\tau_t \ge 2$ なるとなるように更新ごとに射影した．

各実験では，相対累積予測誤差を性能の指標とした．ベンチマークとしては，人工データでは真の生成分布を既知としたときの予測，実データでは学習率 0.1 の AdaGrad による予測を用いた．

人工データの予測　人工データでは，各手法の予測性能の時系列や変化のタイプ (次元，変化期間の長さ，変化の大きさ) に対する依存性を比較した（図 2）．全体を通して vSGD に対して Cogra は予測性能を改善し，よりロバストな予測を得ることができていることがわかる **(Q2)**．また特に，Cogra は変化の程度が大きい時に他手法よりも優れた予測性能を発揮することがわかった **(Q3)**．

実データの予測　最後に，Cogra の実データにおける予測性能を検証した．予測モデルには $p = 3$ の VAR モデルを用いた．用いたデータは ActivityRecognition [Casale 12], EEGEyeState, GasSensorArray [Fonollosa 15] の三種類で，すべて UCI レポジトリ [Lichman 13] から取得した．結果は表 1 の通りで，すべてのデータにおいて Cogra が最高性能を達成しており，次に性能の良い vSGD と比較しても累積予測損失の意味で少なくとも 18%以上改善していることがわかる．



(a) 変化量 μ に対する依存性 ($d = 1$, $T = 0$). (b) 変化期間 T に対する依存性 ($d = 1$, $\mu = 5$). (c) 時系列の次元 d に対する依存性 ($T = 0$, $\mu = 5$).

図 2: **人工データにおける予測性能の依存性.** 各エラーバーは標準偏差を表す.

表 1: **相対累積予測損失.** 各テーブルごとに，第 1 行は SGD の手法，第 2 はハイパーパラメタの設定，第 3 行から第 7 行まではデータセットごとの相対累積予測損失を表す.

	AdaGrad (initial LR)				ADAM (initial LR)				Almeida (initial LR, hyper-LR)		
	0.001	0.01	0.1	1.0	0.001	0.01	0.1	1.0	(10e-9, 10e-3)	(10e-9, 10e-2)	(10e-9, 10e-1)
A(1)	257	3.62	1.0	5.35	9.56	1.95	6.07	138	3.36	3.02	2.01
A(2)	144	2.56	1.0	4.18	5.60	1.76	7.15	125	2.50	2.27	1.63
A(3)	236	4.07	1.0	5.23	9.81	2.06	5.35	68.9	3.29	2.99	2.06
EEG	16.9	0.68	1.0	77.6	2.76	1.02	9.66	945	0.485	0.483	0.482
GAS	8.79	3.10	1.0	48.1	6.15	1.87	3.98	287	1.06e+85	9.82e+65	1.73e+20

	RMSProp (initial LR)					vSGD	Cogra
	0.0001	0.001	0.01	0.1	1.0		
A(1)	80.2	7.76	3.31	200	2.00e+4	0.79	**0.621**
A(2)	46.4	4.60	2.32	110	1.10e+4	0.85	**0.701**
A(3)	74.4	7.84	2.55	120	1.20e+4	0.76	**0.595**
EEG	20.8	2.13	31.6	3.13e+3	3.13e+5	0.30	**0.182**
GAS	9.16	5.39	7.92	695	6.94e+4	0.84	**0.267**

6. 結論と今後の課題

本稿では非定常時系列のオンライン予測問題を扱った．vSGD において提案されたように，この問題は確率的勾配のモーメント推定問題に帰着される．本研究ではヒューリスティックに設計された vSGD のモーメント推定がバイアスをもち，予測性能に関しても悪影響を与えること発見した．これを解決するために，本研究では SMT と呼ばれるモーメント推定量を提案し，vSGD のモーメント推定を SMT で置き換えることで，ハイパーパラメタなしで変化に適応する SGD，Cogra を得た．実験では，Cogra は vSGD を含む既存手法の予測性能を全てのケースで改善し，特に変化量が大きい時系列においてより性能が優れることがわかった．

今後の方向性としては，Cogra をより VAR 以外の複雑な目的関数に対して応用すること，および強化学習など，環境の変化だけでなく，パラメタの変化に応じても目的関数が変化するような問題設定に拡張することがあげられる．

参考文献

[Almeida 99] Almeida, L. B., Langlois, T., Amaral, J. D., and Plakhov, A.: Parameter Adaptation in Stochastic Optimization, in *On-Line Learning in Neural Networks*, pp. 111–134, Cambridge University Press (1999)

[Casale 12] Casale, P., Pujol, O., and Radeva, P.: Personalization and user verification in wearable systems using biometric walking patterns, *Personal and Ubiquitous Computing*, Vol. 16, No. 5, pp. 563–580 (2012)

[Duchi 11] Duchi, J., Hazan, E., and Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization, *Journal of Machine Learning Research*, Vol. 12, pp. 2121–2159 (2011)

[Fonollosa 15] Fonollosa, J., Sheik, S., Huerta, R., and Marco, S.: Reservoir computing compensates slow response of chemosensor arrays exposed to fast varying gas concentrations in continuous monitoring, *Sensors and Actuators B: Chemical*, Vol. 215, pp. 618–629 (2015)

[Kingma 14] Kingma, D. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014)

[Lichman 13] Lichman, M.: UCI Machine Learning Repository (2013)

[Lütkepohl 05] Lütkepohl, H.: *New Introduction to Multiple Time Series Analysis*, Part I, Springer Berlin Heidelberg (2005)

[Schaul 13] Schaul, T., Zhang, S., and LeCun, Y.: No more pesky learning rates, in *Proceedings of the 30th International Conference on Machine Learning* (2013)

[Tieleman 12] Tieleman, T. and Hinton, G.: Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude, COURSERA: Neural Networks for Machine Learning (2012)