Proceedings of the World Congress on Engineering 2011 Vol II
WCE 2011, July 6 - 8, 2011, London, U.K.

On an Adaptive Filter based on Simultaneous Perturbation Stochastic Approximation Method

Hong Son Hoang, Rémy Baraille

***Abstract*—In this paper, the simultaneous perturbation stochastic approximation (SPSA) algorithm is used for seeking optimal parameters in an adaptive filter developed for assimilating observations in very high dimensional dynamical systems. It is shown that the SPSA can achieve high performance similar to that produced by classical optimization algorithms, with better performance for non-linear filtering problems as more and more observations are assimilated. The advantage of the SPSA is that at each iteration it requires only two measurements of the objective function to approximate the gradient vector regardless of dimension of the control vector. This technique offers promising perspectives for future developement of optimal assimilation systems encountered in the field of data assimilation in meteorology and oceanography.**

***Index Terms*—adaptive filter, minimum prediction error, Schur vector, stability, stochastic approximation**

I. Introduction

DATA assimilation is a technique of (optimally) combining numerical model with observations. Let us first consider the standard 4dVar assimilation algorithm which is formulated as follows [1]: Find the initial state $x(0) := x(t_0)$ which minimizes the objective function

$$J[x(0)] = (1/2)(x^b(0) - x(0))^T B(0)^{-1}(x^b(0) - x(0)) \quad (1)$$
$$+(1/2)\sum_{k=1}^{N}[z(k) - Hx(k)]^T R^{-1}[z(k) - Hx(k)]$$

under the constraint

$$x(k+1) = \Phi x(k) + w(k), k = 0, ... \quad (2)$$

$$z(k) = Hx(k) + \epsilon(k), k = 1, ... \quad (3)$$

here $x(k)$ is the n-dimensional system state at $k := t_k$, Φ is the (nxn) fundamental matrix, $z(k)$ is the p-dimensional observation vector, H is the (pxn) observation matrix, w, ϵ are the model and observation noises. We assume $w(k), \epsilon(k)$ are uncorrelated sequences of zero mean and time-invariant covariance Q and R respectively.

Applying a gradient descent algorithm, at each iteration the gradient $\nabla_\theta J, \theta := x(0)$ is used to determine the direction to search J_{min}. The gradient computation requires a forward integration of the numerical model over the period $[t_1, t_N]$ and backward integration of the adjoint Φ^T of Φ. The development of a discrete adjoint solver for partial differential equations by hand differentiation requires a long development time and it involves the errors resulting from necessary approximations used during the differentiation.

Manuscript received March 23 2011; revised April 06 2011.

H.S. Hoang is with the Service Hydrographique et Ocanographique de la Marine (SHOM), 42 av Gaspard Coriolis 31057 TOULOUSE FRANCE, emails: hhoang@shom.fr.

Rémy Baraille is with the Service Hydrographique et Ocanographique de la Marine (SHOM), 42 av Gaspard Coriolis 31057 TOULOUSE FRANCE, Email: remy.baraille@shom.fr

Another approach for data assimilation is known as sequential. Consider (3) and the interval $[t_{k-1}, t_k]$. Then $z = z(k)$ and $\hat{x}(k/k-1) := x^b(k)$ represents the predicted estimate for $x(k)$. Taking the derivative of $J(\theta)$ and equal it to zero leads to the equation for finding the optimal filtered (or analysis) estimate $\hat{x}(k)$. One has now

$$\hat{x}(k) = \hat{x}(k/k-1) + K(k)[z(k) - H\hat{x}(k/k-1)] \quad (4)$$

$$K(k) = B(k)H^T[HB(k)H^T + R]^{-1} \quad (5)$$

in which $K(k)$ represents the filter gain, $\zeta(k) = z(k) - H\hat{x}(k/k-1)$ is the innovation vector, $B(k) := M(k)$ represents the covariance of the prediction error (PE) $e(k/k-1) := \hat{x}(k/k-1) - x(k)$. The unbiased minimum variance (MV) estimate for $x(k)$ is obtained from the Kalman filter (KF) (see [2]). The closed system of equations for the KF includes Algebraic Riccati Equation (ARE). However solving the ARE for the system with state dimensions $10^{12} - 10^{14}$ is impossible.

To deal with these difficulties in [3] the filter is assumed to be of the form (4) with the gain $K(k) := K(k;\theta)$ being given up to a vector of unknown parameters θ. The optimal filter is obtained by minimizing the prediction error (PE) for the system output

$$J(\theta) = E[\Psi(\zeta(k))] \rightarrow \min_\theta, \Psi[\zeta(k)] = ||\zeta(k)||^2 \quad (6)$$

where $||.||$ denotes the l_2 norm. As the use of estimated parameters $\hat{\theta}(k)$ can deviate the filter from its stable behavior, in [5]-[6] the gain is proposed to be selected in order to ensure a filter stability, independently on whatever are the values of tuning gain parameters.

The purpose of this paper is to explore what potential benefits may be achieved by using SPSA algorithm [8]. The essential feature of SPSA is its underlying gradient approximation that requires only two measurements of objective function regardless of the dimension of θ. This feature allows for a significant decrease of the cost of optimization, especially without development of the adjoint code for the tangent linear model (TLM) of the system dynamics.

II. SPSA algorithm

A. Stochastic approximation [4]

Consider the problem of minimizing the objective function

$$\text{Find } \theta^* \text{that solves } min_\theta J(\theta) \quad (7)$$

For an unconstrained optimization, many iterative algorithms rely on the gradient vector $g(.)$ of the objective

ISBN: 978-988-19251-4-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

function. The stochastic approximation (SA) algorithm has the form

$$\theta_{k+1} = \theta_k - a_k Y(\theta_k) \quad (8)$$

where $Y_k = g(\theta_k) + noise$, a_k is a non-negative gain sequence that must satisfy certain conditions [8].

When only the measurements of the objective function are available, $y_k = J(\theta_k) + noise$, one-sided or two-sided gradient approximations, i.e. $g_{ki}(\theta_k) = \frac{y(\theta_k + c_k e_i) - y(\theta_k)}{c_k}$ or $g_{ki}(\theta_k) = \frac{y(\theta_k + c_k e_i) - y(\theta_k - c_k e_i)}{2c_k}$ are of common use where e_i denotes a vector with a one in the ith place and zeros elsewhere. These approximations require $n_\theta + 1$ (or $2n_\theta$, n_θ is the dimension of θ) integrations of the numerical model. For optimization problems with very large n_θ, such algorithms are expensive and in general they are inappropriate for solving assimilation problems.

B. Simultaneous Perturbation Stochastic Approximation (SPSA)

The difficulties due to high dimension of θ can be overcome by applying the SPSA algorithm [8]. In such algorithm, all elements of θ_k are randomly perturbed together to obtain two measurements $y(.)$, but each component of $g_k(\theta_k)$ is formed from a ratio involving the individual components in the perturbation vector and the difference in the two corresponding measurements. For two-sided SP, we have $g_{ki}(\theta_k) = \frac{y(\theta_k + c_k \Delta_k) - y(\theta_k - c_k \Delta_k)}{2c_k \Delta_{ki}}$ where Δ_{ki} can be chosen as the random variable having the symmetric Bernoulli (+/-) 1 distribution. Two common distributions that do not satisfy the conditions for Δ_{ki} are the uniform and the normal.

III. Adaptive filter

A. Adaptive filter

Consider the KF (4)(5) for solving a filtering problem in the system (2)(3). Then $B(k) := M(k)$ satisfies the Riccati equation

$$M(k) = \Phi(k) P(k-1) \Phi^T(k) + Q, P(k) = [I - K(k)H] M(k) \quad (9)$$

Due to very expensive computational burden in time stepping the prediction error covariance matrix (ECM) $M(k)$ in (9), the KF is impractical for solving data assimilation problems. For suboptimal KFs for data assimilation, see [1],[9]. One of possible ways to overcome these difficulties is to choose a parametrized structure of the filter gain from some criteria. In [5] this question is studied from the point of view of the filter stability. The following time-invariant structure of the gain is proposed

$$K = P_r \Theta K_e, K_e = H_e^T [H_e H_e^T + R]^{-1}, H_e = H P_r, \quad (10)$$
$$\Theta = \text{ diag } [\theta_1, ..., \theta_{n_e}], \theta_l \in (0, 2)$$

where $K_e : R^p \to R^{n_e}$ represents the gain mapping the innovation vector from the observational space into the reduced space R^{n_e} of dimension $n_e \leq n$; P_r is mapping from the reduced space R^{n_e} to the full space R^n. The choice of the reduced space plays an important role in assuring a stability of the filter. As proved in [5], under detectability condition, stability of the filter is ensured by forming the columns of P_r from a subspace spanned by leading eigenvectors (or Schur vectors) of the fundamental matrix Φ. The AF is obtained by minimizing the objective function (6). In the AF the gain K in (9) becomes time-varying.

IV. Numerical procedure for construction of the projection subspace

A. Computation of projection subspace spanned by leading Schur vectors [10]

The idea for construction of P_r based on stability criteria is outlined as follows. Let L be an integer number satisfying $1 \leq L \leq n$. Given an $n \times L$ matrix X_0 with orthonormal columns, the method of orthogonal iteration generates the sequence of matrices X_i,

$$S_i = \Phi X_{i-1}, X_i G_i = S_i, i = 1, 2, ...,$$

where X_i is orthonormal. Thus, at i iteration, the columns of X_i are orthonormal vectors which are derived by : 1) integration of the model from each column of X_{i-1} to produce S_i; 2) orthonormalization of L columns of S_i. One sees that the columns of S_i belong to the space spanned by the vectors of X_i.

Let

$$X^T \Phi X = T = \text{diag}(\lambda_i) + \bar{N}, |\lambda_1| \geq |\lambda_2| \geq ... \geq |\lambda_n| \quad (11)$$

be a real Schur decomposition, $X = [X^1, X^2]$, X^1 is $n \times L$ submatrix, $\bar{N}$ is a block upper triangular. The blocks on the diagonal of $\bar{N}$ are of size 1x1 (in which case they represent real eigenvalues) or 2x2 (in which case they are derived from complex conjugate eigenvalue pairs). As shown in [10] (section 7.3), under mild conditions, the distance between $D_L(\Phi) := R[X^1]$ and $R[X_i]$ is of order $O(|\frac{\lambda_{L+1}}{\lambda_L}|^i)$ where $R[X_i]$ denotes the linear space spanned by the columns of X_i. Considering the columns of X_i as patterns for DScVs, the orthogonal iteration method allows us to generate patterns for DScVs which can be used to perform the operator P_r. This method constitutes a basis of the procedure described in the next subsection for generating PE samples which serve as a source for estimating the statistics of PE and to approximate the filter gain (called a PEF - Prediction Error Filter).

B. Procedure for generating dominant prediction error (DPE) samples

Suppose that at the moment $i := t_i$ we are given $x_f(i)$ - some estimate for the true system state $x(i)$. The estimation error is $\delta x(i) = x_f(i) - x(i)$. Integration of the model from $x_f(i)$ produces the prediction $x_p(i+1) = \Phi x_f(i)$ at $t_{i+1} = t_i + \delta t$. Here Φ represents model integration over the interval δt. The true system state at t_{i+1} is $x(i+1) = \Phi x(i)$ (for no model error case). We have then the PE $e_p(i+1) = \Phi x_f(i) - \Phi x(i) = \Phi \delta x(i)$. Thus integrating $\delta x(i)$ by the model Φ yields the vector $s_{t+1} = \Phi \delta x(i)$ which can be

ISBN: 978-988-19251-4-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

considered as a PE pattern growing over the period of model integration. If we apply this procedure to an ensemble of orthogonal columns $X_i = [\delta x^1(i), ..., \delta x^L(i)]$ instead of one vector $\delta x(i)$, the iteration $S_i = \Phi X_{i-1}, X_i G_i = S_i, i = 0, 1, ...$ (see section 3.1) will produce the sequence $\{X_i\}$ approaching L dominant Schur vectors of Φ. The relation $X_i G_i = S_i$ guarantees that the columns of S_i belong to the space spanned by the columns of X_i hence they will approach a subspace spanned by DScVs. As the columns of S_i at the same time represent the PE samples, they will be referred to in the future as *DPE patterns* (in the DScVs subspace). These columns are thus not the patterns randomly generated as done in the EnBF [11],[12] but selected to be "representative" (in a DScV subspace) for the PE. These DPE samples will be used in this paper to estimate the elements or parameters of the ECM which plays an essential role in computation of the filter gain K. We summarize this DPESP for generating the ensemble of PE patterns in the L-dimensional DScVs subspace as follows:

Suppose we want to simulate T patterns for each of the first L Schur vectors of the system dynamics Φ. At $i = 0$, let $x_f(i)$ be an initial estimate for $x(i)$. Suppose we are given the orthogonal matrix X_i, $X_i^T X_i = I_L$ whose columns are L orthonormal perturbations $\delta x_f^l(i), l = 1, ..., L$.

Step 1. For $i \leq T$: Let $x_f(i)$ and X_i be given. Integrate the model $L + 1$ times for producing $x_p(i+1) = \Phi(x_f(i))$ and $x_p^{'l}(i+1) = \Phi(x_f(i) + \delta x_f^l(i)), l = 1, ..., L$. The new matrix $S_{i+1}(L) := [\delta x_p^1(i+1), ..., \delta x_p^L(i+1)]$ is performed whose columns are

$$\delta x_p^l(i+1) = \Phi(x_f(i) + \delta x_f^l(i)) - \Phi(x_f(i)), l = 1, ..., L$$

Step 2. Apply the Gram-Schmidt orthogonalization procedure (see [10]) $X_{i+1} G_{i+1} = S_{i+1}$ to S_{i+1}. The resulting orthonormal perturbations $\{\delta x_f^l(i+1), l = 1, ..., L\}$ are the columns of the matrix $X_{i+1} = [\delta x_f^1(i+1), ..., \delta x_f^L(i+1)]$.

Step 3. If $i + 1 > T$: Stop the procedure. Otherwise set $i := i + 1$ and go to Step 1 subject to $x_f(i+1)$ and X_{i+1}.

Comment 4.1. The DPESP algorithm can be applied to a nonlinear system dynamics where $F(x)$ stays instead of Φx with the modification $\Phi \delta x(i) \approx F[x(i) + \delta x(i)] - F[x(i)]$. The columns of X_i then tend to DScVs of the TLM.

V. Estimation of parameters in periodic signals

A. Numerical model. Assimilation problem

In the observation model

$$z(t_k) = f(t_k) + \epsilon(t_k), f(t) = \sum_{i=1}^{N} \alpha_i cos(\omega_i t), k = 1, ..., M \quad (12)$$

let N be given and $M > 2N$, $\epsilon(t_k)$ is a sequence of zero mean and variance σ^2. The experiment is performed for estimating the parameters α_i, ω_i (see [13]) where $\epsilon(.)$ represents the observation error.

Following [13], for $u_i(t) := \alpha_i cos(\omega_i t)$, we have

$$\frac{d^2 u_i}{dt^2} = -\omega_i^2 u_i, u_i(0) = \alpha_i, \frac{du_i}{dt}(0) = 0 \quad (13)$$
$$z(t_k) = \sum_{i=1}^{N} u_i(t_k), k = 1, 2, ...$$

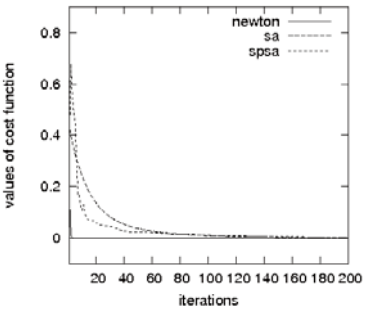

Fig. 1. The values of the objective functions resulting from three algorithms: Newton, SA and SPSA

The vector of unknown parameters $\theta = (\alpha_1, ..., \alpha_N, \omega_1, ..., \omega_N)$ will be estimated by solving the variational problem (for stationary $\epsilon(t_k)$)

$$J(\theta) \rightarrow min_\theta, J(\theta) = \frac{1}{M} \sum_{k=1}^{M} [z(t_k) - f(t_k; \theta))^2] \quad (14)$$

The system (13) can be reformulated in the form equivalent to (2)(3) as

$$\frac{dx(t)}{dt} = 0, x(0) = \theta \quad (15)$$
$$z(t_k) = h[x(t)] = \sum_{i=1}^{N} \alpha_i cos(\omega_i t), k = 1, ..., M$$

The task is to estimate the initial system state $x(0)$ by minimizing (14) using the available observations. Comparing (14) with (1) shows that the first term in the right hand side of (1) does not participate in (14) which is equivalent to assuming that there is no a priori information on θ. The observation operator in the present case is non-linear.

B. Numerical results

The true values of the parameters (α_i, ω_i) are that given in [13], i.e. $\alpha_1 = 1, \alpha_2 = 0.5, \alpha_3 = 0.1$, $\omega_1 = 1.11, \omega_2 = 2.03, \omega_3 = 3.42$. The corresponding initial values are: $\alpha_1(0) = 0.9, \alpha_2(0) = 0.6, \alpha_3(0) = 0.2$; $\omega_1(0) = 1, \omega_2(0) = 2, \omega_3(0) = 3$; $\alpha_1(0) = 0.9, \alpha_2(0) = 0.6, \alpha_3(0) = 0.2$. The measurements are contaminated by the Gaussian noise $N(./m, \sigma^2)$ with $m = 0$ and $\sigma^2 = 0.01$. To estimate θ we will apply the well known Newton iterative algorithm (see [10]) and two versions of the stochastic optimization algorithm: SA (8) and SPSA (section 1, B).

From Fig. 1 one sees that the objective function, associated with the Newton-algorithm, decreases most quickly. Compared to the SA, the SPSA-algorithm works better in minimizing the objective function. Globally, all three algorithms are capable of well tuning the parameters to decrease the objective function.

Fig. 2 shows the consistency of three algorithms in estimating α. The SPSA-algorithm produces more efficient estimate for α compared to the SA-algorithm. As to the estimates for ω, Fig. 3 demonstrates that with noisy measurements, the strategy to fit fast and exactly the output of the model to measurements can lead to big biased estimate as it happens in the Newton-algorithm.

ISBN: 978-988-19251-4-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

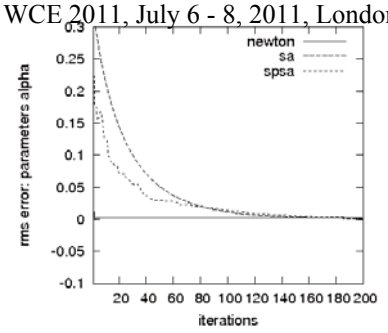

Fig. 2. Consistency of three algorithms in estimating α

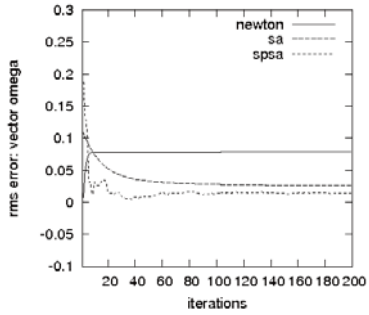

Fig. 3. Estimation errors for ω in three algorithms

VI. Algorithm of the ROAF for altimetric SSH data assimilation

A. MICOM model and observations

The Miami Isopycnal Coordinate Ocean Model (MICOM), used here for the twin experiment is identical to that described in [7]. The model configuration is a domain situated in the North Atlantic from 30^0 N to 60^0 N and 80^0 W to 44^0 W; for the exact model domain and some main features of the oceanic current (mean, variability of the sea surface height (SSH), velocity ...) produced by the model, see [7]. The observations are SSH taken from the control run every 10 days (ds), only at the grid points $i_o = 1, 11, ..., 131$, $j_o = 1, 11, ..., 171$ from the grid $i = 1, 140; j = 1, 180$. They are noise-free.

B. Reduced-order filter and gain structures

The filter used for assimilating SSH observations is of the form

$$\hat{x}(k) = F[\hat{x}(k-1)] + KP_{oi}\zeta(k), k = 0, 1, ... \quad (16)$$

where $\hat{x}(k)$ is the filtered estimate for $x(k)$, $x(k) = [h(k), u(k), v(k)]$ is the system state at $k := t_k$, $t_{k+1} - t_k = 10\ ds$, $F(.)$ represents integration of the MICOM nonlinear model over 10 ds, K is the filter gain, $\zeta(k)$ is the innovation vector. The operator P_{oi} will interpolate the missing SSH from observed points. The gain K is symbolically given by $K = (K_h, K_u, K_v)^T$ with K_u, K_v representing the operators which produce the correction for the velocity (u, v) from the layer thickness correction $K_h P_{oi}\zeta(k)$ using the geostrophy hypothesis. As SSH observations are linear functions with respect to h, the observation equation is given by (3) (see [7]). By considering $P_{oi}z$ instead of z, the observation operator H is of the form

$$H = [I_p, ..., I_p] \quad (17)$$

where I_p is the unit matrix of dimension pxp ($p = N_h$ is the number of all horizontal grid points).

C. Structure of the ECM for PE and its estimation

The ECM $M(k)$ is assumed to be constant and of the form

$$M(k) = \Omega = [\omega_{l,m}]_{l,m=1}^{N_z} \otimes I_p, \quad (18)$$

where $\otimes$ denotes the Kronecker product; N_z is the number of thickness layers in the model, ω_{lm} is a scalar representing the covariance of the PE between two layers l and m. The elements ω_{lm} can be chosen a priori from physical considerations or estimated from error patterns. In the Cooper-Haines filter (CHF, see [14], [7]), the elements ω_{lm} are deduced from several physical constraints like conservation of potential vorticity, no motion at the bottom layer ... In the PEF ω_{lm} are estimated using the patterns of DScVs. Applying the DPESP subject to $L = 1$ yields the ensemble of DPE patterns $\delta h_p(i, j, lr; k), k = 1, ..., T$ from which one estimates ω_{lm} by

$$\omega_{lm}(T) = \frac{1}{T}\sum_{k=1}^{T} \mu_{l,m}^k, \quad (19)$$

$$\mu_{l,m}^k = \frac{1}{p}\sum_{i,j} \delta h_p(i, j, l; k)\delta h_p(i, j, m; k)$$

where i, j span all horizontal grid points whose number is equal to p. The terms $\frac{1}{T}, \frac{1}{p}$ should be replaced by $\frac{1}{T-1}, \frac{1}{p-1}$ for $T > 1$, $p > 1$ to provide the unbiasedness of the estimates. As the ensemble $\delta h_p(i, j, lr; k), k = 1, ..., T$ is generated by the model alone, for fixed T, the matrix Ω is constant.

We will apply the SA algorithms for seeking the (sub)optimal filters in two class of parametrized filters based on : 1) the CHF and 2) the PEF. The difference between the PEF and CHF is lying in the way we estimate the elements of Ω. Substituting Ω from (19) into (5) and for $R = \sigma_r^2 I_p$ leads to

$$K_h = [k(1)I_p, ..., k(N_z)I_p]^T, \quad (20)$$

$$k(l) = \sum_{m=1}^{N_z} \frac{\omega_{l,m}}{s}, s = \sum_{m,m'=1}^{N_z} \omega_{m,m'} + \sigma_r^2$$

hence $k(l)$ is a scalar, $l = 1, ..., N_z$. The Cooper-Haines filter (CHF) [14] is obtained from (20) under hypotheses [7] on the conservation of linear potential vorticity and of no correction for the velocity at the bottom layer. For the noise-free observations, the parametrized gain in the CHF is of the form [7]

$$K_{chf} = [(1-\theta_2\alpha), (\theta_2-\theta_3)\alpha, (\theta_3-\theta_4\alpha), \theta_4\alpha]^T \otimes I_p \quad (21)$$

For the present MICOM model, $\alpha = -184.965$. The CHF in [14] corresponds to $\theta_l = 1, l = 2, 3, 4$ and has the form

$$K_{chf} = [185.965, 0, 0, -184.965I_p]^T \otimes I_p \quad (22)$$

ISBN: 978-988-19251-4-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

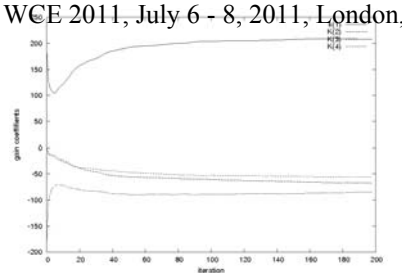

Fig. 4. Estimated gain coeffficients as functions of iteration. One sees a quick convergence of the gain coefficients.

D. Parametrization of the gain for the PEF. Adaptive filter

Following the ROAF approach based on the gain structure (10), the Cholesky decomposition method is used to decompose $M(k) = \Omega$ as

$$\Omega = DD^T \tag{23}$$

Subject to (23), the gain (10) is equal to

$$K = P_r \Theta K_e, P_r = D, \theta_l \in (0, 2) \tag{24}$$

with K_e defined as in (10). In the adaptive filter, the diagonal elements of Θ are adjusted to minimize the prediction error for the SSH variable.

As the coefficient ω_{lm} represents the covariance of the PE between two layers l and m, they can be estimated using the simulated DPE patterns obtained from the DPESP. In the experiment to follow we will generate the ensemble of T patterns $\delta h_p(i, j, lr; k), k = 1, ..., T$ by applying DPESP subject to $L = 1$. The elements ω_{lm} are estimated by (19).

In the adaptive PEF (APEF), the gain (23)(24) is parametrized with

$$\Theta = \text{ diag } [\theta_1, \theta_2, \theta_3, \theta_4] \otimes I_p, \theta_l \in (0, 2), l = 1, l = 1, ..., 4$$

The initial values $\theta_l(0) = 1, l = 1, ..., 4$ correspond to the non-adaptive PEF. For the noisy-free observations, $R = 0$, this leads to the gain

$$K_{pef} = [205.506, -62.919, -58.478, -83.107]^T \otimes I_p \tag{25}$$

Figure 4 shows the gain coefficients computed in accordance with (20) which are functions of iteration T. Compared with the gain in the CHF (22) one sees that the gains in two filters CHF and PEF are of nearly the same magnitude for the 1st layer but the physical hypotheses (**H2**), (**H3**) ignore the correction to be made for the intermediate layers $l = 2, 3$. In the PEF these corrections remain important to maintain the better performance of the PEF (see next sections).

E. Adaptive algorithms for the CHF and PEF

Consider two sets of filters with the gain (21) and (23),(24),(25). The adaptive versions for the CHF and PEF (denoted as ACHF and APEF) are obtained by varying the vector of parameters θ to minimize the mean of the SSH prediction error. Let the initial values for θ be $\theta_l = \theta_l(0) = 1, l = 1, 2, 3, 4$ which correspond to the non-adaptive CHF and PEF.

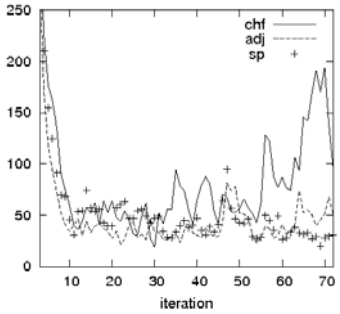

Fig. 5. Sample objective functions resulting from three filters CHF, ACHF(SP), ACHF(ADJ)

TABLE I
ERROR REDUCTION (IN PERCENTAGE) ACHIEVED BY ACHF(SP) AND ACHF(ADJ)

	ER1(%)	ER2(%)
J	22,1	26,2
$e_u(p)$	12,9	18,8
$e_u(f)$	15,4	20,4
$e_v(p)$	15,6	19,6
$e_v(p)$	14,5	18,6
$e_{uv}(p)$	16,7	21,1
$e_{uv}(f)$	15	19,6

VII. NUMERICAL RESULTS

A. Adaptive CHF

In Table I the the error reductions by ACHF(ADJ) (using gradient measurements computed by adjoint equation) and by ACHF(SP) (SPSA using measurements of cost function) are displayed where ER1(%), ER2(%), expressed in percentage, show how the corresponding ACHF(SP) or ACHF(ADJ) has reduced rms (root-mean square) of estimation errors compared to that of CHF. For example, over the large window $k \in [5 : 72]$, the SPSA algorithm has reduced about 15 % rms errors whereas this percentage is of order 20 % if the gradient is computed by the adjoint code. Figure 5 depicts instantaneous values of the objective function resulting from three filters.

To see in detail what happens really during the period of last four months, i.e. $k \in [61 : 72]$, Table II displays the RMS-PE and RMS-FE resulting from three filters. As expected, the ACHF(SP) behaves now better than the ACHF(ADJ), with the reduction of velocity error by more than 10 %. As to the CHF, during this period one observes an important increase of estimation error (see Fig. 5).

TABLE II
RMS OF ESTIMATION ERRORS AVERAGED OVER $k \in [61 : 72]$

Filter	CHF	ACHF(SP)	ACHF(ADJ)
$J(cm)$	11.53	5.48	7.02
$e_u(p)(cm/s)$	9.25	5.03	5.88
$e_u(f)(cm/s)$	7.57	4.49	5.04
$e_v(p)(cm/s)$	9.36	5.28	6.15
$e_v(f)(cm/s)$	8.05	4.72	5.28
$e_{uv}(p)(cm/s)$	8.94	4.96	5.78
$e_{uv}(f)(cm/s)$	7.51	4.43	4.96

ISBN: 978-988-19251-4-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

RMS OF ESTIMATION ERRORS AVERAGED OVER $k \in [5:72]$

Filter	PEF	APEF(SP)	APEF(ADJ)	ER1 (%)	ER2 (%)
J	6.36	5.90	5.88	7.2	7.5
$e_u(p)$	5.69	5.42	5.34	4.7	6.2
$e_u(f)$	4.77	4.48	4.45	6.1	6.7
$e_v(p)$	5.74	5.43	5.36	5.4	6.6
$e_v(f)$	5.10	4.83	4.79	5.3	6.1
$e_{uv}(p)$	5.57	5.24	5.20	5.9	6.6
$e_{uv}(f)$	4.90	4.62	4.59	5.7	6.3

TABLE IV
RMS OF ESTIMATION ERRORS AVERAGED OVER $k \in [61:72]$

Filter	PEF	APEF(SP)	APEF(ADJ)	ER1(%)	ER2(%)
J	6.94	5.75	5.96	17.1	14.1
$e_u(p)$	5.99	5.04	5.23	15.9	12.7
$e_u(f)$	5.26	4.44	4.59	15.6	12.7
$e_v(p)$	6.19	5.19	5.41	16.2	12.6
$e_v(f)$	5.47	4.56	4.82	18.5	11.9
$e_{uv}(p)$	5.85	4.92	5.11	15.9	12.6
$e_{uv}(f)$	5.15	4.32	4.52	16.1	12.2

B. Adaptive PEF

Table III-IV show that the PEF is much more efficient than the CHF and it slightly outperforms the ACHF(SP) and ACHF(ADJ). Thus the statistics extracted from DPE samples play the important role in correct estimating the filter gain and in improving the filter performance.

As the errors in the PEF are much lower than those produced by the CHF, there remains no great margin for reducing the errors in the PEF by adaptation compared to the case of optimizing the CHF structure. Even though, as seen in Tables III-IV, the adaptation remains still as advantageous tool for improving the performance of the PEF. For the assimilation period, compared with the PEF, the adaptation allows to reduce the rms estimation error by about 5-6 % in the APEF(SP) and 6-7 % in the APEF(ADJ). These reductions are less important than that achieved by the ACHF(SP) and ACHF(ADJ) with respect to the CHF (they are equal to 15 % and 20 % respectively, see Table I). At the last 4 months of assimilation, the APEF(SP) again outperforms the APEF(ADJ). Meantime, the error reduction is achieved by 16-17 % in the APEF(SP) and by 12-13 % in the APEF(ADJ) compared to the non-adaptive PEF. The performance of the APEF presented here is based on the gain parametrization consisting of each parameter for each layer thickness. Due to space limit of this paper we cannot present here the way to parametrize the gain in 3d space. In this situation the number of parameters to be updated in the gain is equal to 140x180x4 = 100800 elements. By this way one can reduce more efficiently the filtered errors at the same computational cost as shown in this paper for 4 parameters since the SPSA uses simultaneous perturbations to approximate the gradient vector. We do hope to present these interesting results in an expanded version of this paper.

VIII. Conclusions

The objective of this paper is to present a very simple tool named as SPSA for optimization problems in very high dimensional systems and to demonstrate its high efficiency in state-parameter estimation problems which are typically encountered in the field of data assimilation in meteorology and oceanography. The SPSA algorithm is very simple to implement since it requires only two integrations of direct numerical model for estimating the gradient of objective function. For meteorological and oceanic models with dimension of order $10^7 - 10^8$, this method represents a great advantage for future development of optimal assimilation systems. As seen from the numerical experiments, due to a random simultaneous perturbation of all parameters, the SPSA requires more iterations, compared to the adjoint method, to well determine descent direction and to minimize the objective function. On the other hand, the SPSA method seems to be more efficient as iteration progresses, especially in optimizing non-linear systems since it calculates derivatives using the difference between two non-linear integrations of the model whereas the adjoint method approximates the gradient by linearization technique. That is why we found in all experiments the better performance of the AF based on SPSA at the end of assimilation period, compared to that based on the adjoint method.

References

[1] M. Ghil and P. Manalotte-Rizzoli, "Data assimilation in meteorology and oceanography". *Adv. Geophys*, 33, pp. 141-266, 1991.
[2] A.E. Bryson and Y.C. Ho, *Applied optimal control*. Washington, DC: Hemisphere, 1975.
[3] H.S. Hoang, P. De Mey, O. Talagrand and R. Baraille, "A new reduced-order adaptive filter for state estimation in high dimensional systems," *Automatica*, 33, pp. 1475-1498, 1997.
[4] YA. Zypkin, *Adaptation and Learning in Automatic Systems*, New York, Academic, 1971.
[5] H.S. Hoang, O. Talagrand and R. Baraille, "On the design of a stable filter for state estimation in high dimensional systems", *Automatica*, 37, pp. 341-359, 2001.
[6] H.S. Hoang, O. Talagrand and R. Baraille, "On the stability of a reduced-order filter based on dominant singular value decomposition of the systems dynamics", *Automatica*, 45, pp. 2400-2405, 2009.
[7] H.S. Hoang, O. Talagrand and R. Baraille, "On an adaptive filter for altimetric data data assimilation and its application to a primitive equation model MICOM", *Tellus*, 57A, no 2, pp. 153-170, 2005.
[8] C.S. Spall, "An Overview of the Simultaneous Perturbation Method for Efficient Optimization", *Johns Hopkins Apl Tech. Digest*, V. 19, No 4, pp. 482-492, 1998.
[9] R. Todling and S.E. Cohn, "Suboptimal schemes for atmospheric data assimilation based on the Kalman filter", *Mon. Wea. Rev.*, 122, pp. 2530-2557, 1994.
[10] G.H. Golub and C.F. Van Loan C.F., *Matrix Computations*, 2 edn. Johns Hopkins, 1993.
[11] T.M. Hamill, "Ensemble-based atmospheric data assimilation", in *Predictability of Weather and Climate*, Cambridge Univ. Press, 2006, pp. 124-156.
[12] G. Evensen G., "The ensemble Kalman filter: Theoretical formulation and practical implementation", *Ocean Dynamics*, 53, pp. 343-367, 2003.
[13] R.E. Bellman and R.E. Kalaba, *Quasi-linearization and non-linear boundary-value problems*, Elsevier, New-York, 1965.
[14] M. Cooper and K. Haines, "Altimetric assimilation with water property conservation", *J. Geophys. Res.*, 101, pp. 1059-1077, 1996.

ISBN: 978-988-19251-4-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)