PROJECTION-BASED DUAL AVERAGING FOR STOCHASTIC SPARSE OPTIMIZATION

Asahi Ushio[†]*, Masahiro Yukawa*[†]

[†] Department of Electronics and Electrical Engineering, Keio University, Japan

ABSTRACT

We present a variant of the regularized dual averaging (RDA) algorithm for stochastic sparse optimization. Our approach differs from the previous studies of RDA in two respects. First, a sparsity-promoting metric is employed, originated from the proportionate-type adaptive filtering algorithms. Second, the squared-distance function to a closed convex set is employed as a part of the objective functions. In the particular application of online regression, the squared-distance function is reduced to a normalized version of the typical squared-error (least square) function. The two differences yield a better sparsity-seeking capability, leading to improved convergence properties. Numerical examples show the advantages of the proposed algorithm over the existing methods including ADAGRAD and adaptive proximal forward-backward splitting (APFBS).

Index Terms— online learning, sparse optimization, stochastic optimization, orthogonal projection, proximity operator

1. INTRODUCTION

Stochastic optimization (stochastic approximation [1] more in general) has drawn growing attention over the past years due particularly to the recent data deluge [2]. We focus on the case where the solution is "sparse"; i.e., many components are zero. This often happens in a wide range of applications such as echo cancellation, channel estimation, text classification, etc. Sparseness has been exploited in adaptive filtering [3–5] which is closely related to stochastic optimization. The algorithms in [3–5] can be regarded as variable-metric methods [6, 7]. More recently, sparsity-aware algorithms have been studied for stochastic optimization and online learning, including the adaptive proximal forward-backward splitting (APFBS) method [8, 9], the FOBOS method [10], and the regularized dual averaging (RDA) method [11]. In particular, the idea of RDA comes originally from the primal-dual subgradient methods [12] of Nesterov, and it is known to yield a sparser solution than the FOBOS method [11]. An approach similar to RDA is known as the follow-the-regularized-leader in online convex optimization [13]. The objective of this paper is to improve the performance of RDA by leveraging the insights of sparsity-aware adaptive filtering. The key ingredients of the proposed algorithm are (i) normalization of the input vector and (ii) variable metric.

In adaptive filtering, it is well known that the normalized least mean square (NLMS) algorithm [14, 15] often performs better and is more stable than the popular stochastic gradient descent (SGD) method referred to as the least mean square (LMS) algorithm [16]. The NLMS algorithm is usually derived based the so-called minimum disturbance principle [17], and is widely recognized as an iterative projection method onto a zero-instantaneous-error hyperplane. In the present study, we highlight the fact that NLMS can be regarded as a stochastic gradient method for a "normalized" squared error cost, which is equivalent to the squared distance function to the zero-instantaneous-error hyperplane. The squared distance functions have actually been considered in the studies of the adaptive projected subgradient method (APSM) [18–20] and APFBS.

In this paper, we present a sparse stochastic optimization algorithm called *projection-based dual avearaging (PDA)*. We consider a squared-distance function to a random closed convex set, where the randomness comes from the measurements. To be more specific, we consider a specific stochastic optimization problem of minimizing the expectation of the squared distance function penalized by some convex regularizer. Here, the distance is defined with the variable metric, which is denoted by $\boldsymbol{Q}_t$, that aims to promote sparsity of our estimates. The PDA update involves regularization by the squared $\boldsymbol{Q}_t$-norm of the coefficient vector. PDA differs from the previous studies of RDA in this respect in addition to the difference in the cost functions. (See Section 3.2 for the differences from ADAGRAD-RDA.) As a result, the dual-variable vector is updated with the $\boldsymbol{Q}_t$-gradient of the squared-distance function. This makes two practical advantages in online regression involving sparse structures. First, the use of the squared-distance function avoids the situation that the gradient vector becomes undesirably large for large inputs, stabilizing the algorithm. Second, the use of $\boldsymbol{Q}_t$-metric guides the update direction towards the true solution. Assembling them together, the proposed algorithm enjoys a notable sparsity-seeking property. Numerical examples for sparse-system estimation and echo cancellation show the advantages of the proposed algorithm.

2. PRELIMINARIES

2.1. Projection-based Method

We denote by $\mathbb{R}^{a\times b}$ the set of real $a \times b$ matrices, and by $\mathbb{N}$ the set of all nonnegative integers. Also we denote by $\boldsymbol{w}^\mathsf{T}$ the transpose of a vector $\boldsymbol{w} := [w_1, w_2, \cdots, w_n]^\mathsf{T} \in \mathbb{R}^n$. We consider an online regularized stochastic optimization problem

$$\min_{\boldsymbol{w}\in\mathbb{R}^n} \mathbb{E}\left[\varphi_t(\boldsymbol{w})\right] + \psi_t(\boldsymbol{w}), \quad t \in \mathbb{N}, \tag{1}$$

where t is the time index, $\mathbb{E}$ stands for expectation, ψ_t is a possibly nonsmooth regularization term, and

$$\varphi_t(\boldsymbol{w}) := \frac{1}{2} d^2_{\boldsymbol{Q}_t}(\boldsymbol{w}, C_t). \tag{2}$$

Here, $d_{\boldsymbol{Q}_t}(\boldsymbol{w}, C_t) := \min_{\boldsymbol{z}\in C_t} ||\boldsymbol{w} - \boldsymbol{z}||_{\boldsymbol{Q}_t}$ is the $\boldsymbol{Q}_t$-metric distance for a positive definite matrix $\boldsymbol{Q}_t := \mathrm{diag}(q_{t,1} \ldots, q_{t,n}) \in \mathbb{R}^{n\times n}$ between an arbitrary point $\boldsymbol{w} \in \mathbb{R}^n$ and a closed convex set $C_t \subset \mathbb{R}^n$, where $||\boldsymbol{w}||_{\boldsymbol{Q}_t} := \sqrt{\langle \boldsymbol{w}, \boldsymbol{w}\rangle_{\boldsymbol{Q}_t}}$ is the $\boldsymbol{Q}_t$-norm induced by the inner product $\langle \boldsymbol{w}, \boldsymbol{z}\rangle_{\boldsymbol{Q}_t} := \sqrt{\boldsymbol{w}^\mathsf{T}\boldsymbol{Q}_t\boldsymbol{z}}$, $\boldsymbol{w}, \boldsymbol{z} \in \mathbb{R}^n$. The $\boldsymbol{Q}_t$-gradient of φ_t at the previous estimate $\boldsymbol{w}_{t-1} \in \mathbb{R}^n$ is given by

$$\boldsymbol{g}_t := \nabla_{\boldsymbol{Q}_t}\varphi_t(\boldsymbol{w}_{t-1}) = \boldsymbol{w}_{t-1} - P^{\boldsymbol{Q}_t}_{C_t}(\boldsymbol{w}_{t-1}), \tag{3}$$

This work was partially supported by JSPS Grants-in-Aid (15K06081, 15K13986, 15H02757).

where $P_{C_t}^{\boldsymbol{Q}_t}(\boldsymbol{w}) := \arg\min_{\boldsymbol{z}\in C_t} ||\boldsymbol{w}-\boldsymbol{z}||_{\boldsymbol{Q}_t}$ is the $\boldsymbol{Q}_t$-projection onto C_t. In the case of $\psi_t = 0$ (i.e., the case of unregularized stochastic optimization problems), the SGD update is given by

$$\boldsymbol{w}_t := \boldsymbol{w}_{t-1} - \eta \boldsymbol{g}_t, \tag{4}$$

where $\eta \in [0, 2]$ is the step size. Note here that the projection operator is nonexpansive (i.e., Lipschitz continuous with constant 1), and the gradient operator $\nabla_{\boldsymbol{Q}_t}\varphi_t$ is also nonexpansive. The gradient vector has the following property:

$$\boldsymbol{g}_t = \boldsymbol{0} \Leftrightarrow P_{C_t}^{\boldsymbol{Q}_t}(\boldsymbol{w}_{t-1}) = \boldsymbol{w}_{t-1} \Leftrightarrow \boldsymbol{w}_{t-1} \in C_t. \tag{5}$$

The algorithm (4) is reduced to the proportionate affine projection algorithm (PAPA) [21, 22] if C_t is a set of vectors that makes the instantaneous errors for several previous inputs to be simultaneously zero. If the instantaneous error for the current input is only taken into account, PAPA is reduced to the (improved) proportionate NLMS (PNLMS) algorithms [3, 4, 23]. If the metric is Euclidean, PAPA and PNLMS are further reduced to the affine projection algorithm (APA) [24, 25] and NLMS, respectively.

2.2. Dual Averaging

To solve (1) in the case of $\psi_t = 0$, Nesterov has proposed the dual averaging method in [12], which aims to minimize

$$l_t(\boldsymbol{w}) := \frac{1}{t}\sum_{i=1}^{t}\left[\varphi_i(\boldsymbol{w}_i) + \langle \boldsymbol{g}_i, \boldsymbol{w}-\boldsymbol{w}_i\rangle_{\boldsymbol{I}_n}\right], \tag{6}$$

where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. The lower linear model (6) is an average of affine minorants of $\varphi_i(\boldsymbol{w}_i)$, $i = 1 \dots t$, for $\boldsymbol{Q}_i := \boldsymbol{I}_n$. The simple dual averaging update is given by

$$\begin{aligned}\boldsymbol{w}_t &:= \arg\min_{\boldsymbol{w}\in\mathbb{R}^n}\left(l_t(\boldsymbol{w}) + \mu_t h(\boldsymbol{w})\right)\\ &= \arg\min_{\boldsymbol{w}\in\mathbb{R}^n}\left(\left\langle \frac{\boldsymbol{s}_t}{t}, \boldsymbol{w}\right\rangle_{\boldsymbol{I}_n} + \mu_t h(\boldsymbol{w})\right),\end{aligned} \tag{7}$$

where $\mu_t = \mathcal{O}(\frac{1}{\sqrt{t}})$, and $\boldsymbol{s}_t := \sum_{i=1}^{t}\boldsymbol{g}_i$, $h(\boldsymbol{w})$ is the so-called prox-function. The scheme in (7) has been shown to be primal-dual [12].

3. PROJECTION-BASED DUAL AVERAGING

We present the proposed PDA algorithm, which is a projection-based online sparse estimation framework based on the dual averaging. The update equation is given by

$$\begin{aligned}\boldsymbol{w}_t &:= \arg\min_{\boldsymbol{w}\in\mathbb{R}^n}\left(\langle \boldsymbol{s}_t, \boldsymbol{w}\rangle_{\boldsymbol{Q}_t} + \frac{||\boldsymbol{w}||_{\boldsymbol{Q}_t}^2}{2\eta} + \psi_t(\boldsymbol{w})\right)\\ &= \arg\min_{\boldsymbol{w}\in\mathbb{R}^n}\left(\eta\psi_t(\boldsymbol{w}) + \frac{1}{2}||\boldsymbol{w}+\eta\boldsymbol{s}_t||_{\boldsymbol{Q}_t}^2\right)\\ &= \mathrm{prox}_{\eta\psi_t}^{\boldsymbol{Q}_t}(-\eta\boldsymbol{s}_t),\end{aligned} \tag{8}$$

where the proximity operator is defined, for $\boldsymbol{w} \in \mathbb{R}^n$, as [26, 27]

$$\mathrm{prox}_{\eta\psi_t}^{\boldsymbol{Q}_t}(\boldsymbol{w}) := \arg\min_{\boldsymbol{z}\in\mathbb{R}^n}\left(\eta\psi_t(\boldsymbol{w}) + \frac{1}{2}||\boldsymbol{w}-\boldsymbol{z}||_{\boldsymbol{Q}_t}^2\right). \tag{9}$$

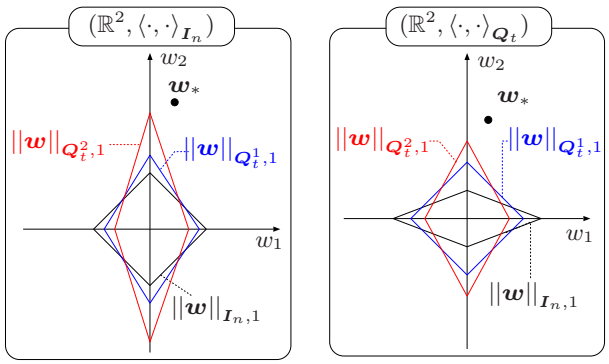


Fig. 1. Unit balls for different norms in $(\mathbb{R}^2, \langle\cdot,\cdot\rangle_{\boldsymbol{Q}_t})$ and $(\mathbb{R}^2, \langle\cdot,\cdot\rangle_{\boldsymbol{I}_n})$ in the case of $q_{t,1} < q_{t,2}$.

3.1. Application to Online Regression

We apply the PDA algorithm to an online regression problem. Let $\boldsymbol{x}_t \in \mathbb{R}^n$ be the input vector, and $y_t := \boldsymbol{w}_*^{\mathsf{T}}\boldsymbol{x}_t + \nu_t \in \mathbb{R}$ is the output at time instant t with the unknown vector $\boldsymbol{w}_* \in \mathbb{R}^n$ and the additive noise $\nu_t \in \mathbb{R}$. Define

$$C_t := \left\{\boldsymbol{w}\in\mathbb{R}^n \mid \left|\left|\boldsymbol{X}^{\mathsf{T}}\boldsymbol{w} - \boldsymbol{y}_t\right|\right|_{\boldsymbol{I}_n}^2 = 0\right\}, \tag{10}$$

where $\boldsymbol{X}_t := [\boldsymbol{x}_t \cdots \boldsymbol{x}_{t-r+1}] \in \mathbb{R}^{n\times r}$ and $\boldsymbol{y}_t := [y_t, \dots, y_{t-r+1}]^{\mathsf{T}} \in \mathbb{R}^r$ for some $r \in \mathbb{N}^* := \mathbb{N}\setminus\{0\}$. The projection onto the linear variety C_t is given by

$$P_{C_t}^{\boldsymbol{Q}_t}(\boldsymbol{w}_{t-1}) := \boldsymbol{w}_{t-1} - \boldsymbol{Q}_t^{-1}\boldsymbol{X}_t^{\dagger}(\boldsymbol{X}_t^{\mathsf{T}}\boldsymbol{w}_{t-1} - \boldsymbol{y}_t), \tag{11}$$

where $\boldsymbol{X}_t^{\dagger}$ is the Moore-Penrose pseudo-inverse. In practice, $\boldsymbol{X}_t^{\dagger}$ is replaced by $\boldsymbol{X}_t(\boldsymbol{X}_t^{\mathsf{T}}\boldsymbol{Q}_t^{-1}\boldsymbol{X}_t + \delta\boldsymbol{I}_n)^{-1}$, where $\delta > 0$ is the regularization parameter for numerical stability. We mention here that $(\boldsymbol{X}_t^{\mathsf{T}}\boldsymbol{Q}_t^{-1}\boldsymbol{X}_t + \delta\boldsymbol{I}_n)^{-1}$ normalizes the input vectors. The metric is designed as follows [28]:

$$\boldsymbol{Q}_t := \frac{\alpha}{n}\boldsymbol{I}_n + \frac{1-\alpha}{S_t}\tilde{\boldsymbol{Q}}_t^{-1}, \tag{12}$$

where $\tilde{\boldsymbol{Q}}_t := \mathrm{diag}(|w_{t-1,1}|, \dots, |w_{t-1,n}|) + \epsilon\boldsymbol{I}_n$ for some $\epsilon > 0$, $\alpha \in [0, 1]$, and $S_t := \sum_{i=1}^{n}(|w_{t-1,i}| + \epsilon)^{-1}$.

The regularization term is defined as

$$\psi_t(\boldsymbol{w}) = \lambda\,||\boldsymbol{w}||_{\boldsymbol{Q}_t^2,1} := \lambda\sum_{i=1}^{n} q_{t,i}^2|w_i|, \tag{13}$$

where $\lambda > 0$ is the regularization parameter. Figure 1 illustrates the unit balls for three norms in the Hilbert spaces $(\mathbb{R}^2, \langle\cdot,\cdot\rangle_{\boldsymbol{I}_n})$ and $(\mathbb{R}^2, \langle\cdot,\cdot\rangle_{\boldsymbol{Q}_t})$: the ℓ_1 norm $||\boldsymbol{w}||_{\boldsymbol{I}_n,1} := \sum_{i=1}^{n}|w_i|$, a weighted ℓ_1 norm $||\boldsymbol{w}||_{\boldsymbol{Q}_t^1,1} := \sum_{i=1}^{n} q_{t,i}|w_i|$, and $||\boldsymbol{w}||_{\boldsymbol{Q}_t^2,1}$ in (13). One can see that the ℓ_1 ball has a "fat" shape in $(\mathbb{R}^2, \langle\cdot,\cdot\rangle_{\boldsymbol{Q}_t})$. This actually forces the proximity operator to shrink the large component more than the small component, yielding an undesirable bias. To avoid this and to shrink the small component more, we employ the norm $||\boldsymbol{w}||_{\boldsymbol{Q}_t^2,1}$ in (13), of which the unit ball has a "tall" shape in $(\mathbb{R}^2, \langle\cdot,\cdot\rangle_{\boldsymbol{Q}_t})$. The proximity operator for the ψ_t in (13) is given by

$$\mathrm{prox}_{\eta\psi_t}^{\boldsymbol{Q}_t}(\boldsymbol{w}) = \sum_{i=1}^{n}\boldsymbol{e}_i\mathrm{sgn}(w_i)\left[|w_i| - q_{t,i}\lambda\eta\right]_+, \tag{14}$$

Table 1. PDA for online regression.

PDA for regression problem
Requirement: $\lambda > 0, \eta \in [0, 2], \alpha \in [0, 1]$ $r \in \mathbb{N}^*, \delta > 0$
Initialization: $\boldsymbol{s}_0 = \boldsymbol{0}$
Iteration: For $t = 0, 1, 2, \cdots$
1. $\boldsymbol{g}_t := \boldsymbol{w}_{t-1} - P_{C_t}^{\boldsymbol{Q}_t}(\boldsymbol{w}_{t-1})$ with (11)
2. $\boldsymbol{s}_t = \boldsymbol{s}_{t-1} + \boldsymbol{g}_t$
3. $\boldsymbol{w}_t := \mathrm{prox}_{\eta\psi_t}^{\boldsymbol{Q}_t}(-\eta \boldsymbol{s}_t)$ with (14)

where $\{\boldsymbol{e}_i\}_{i=1}^n$ is the standard basis of $\mathbb{R}^n$, $\mathrm{sgn}(\cdot)$ is the signum function, and $[\cdot]_+$ is the hinge function. The proposed algorithm for online regression is summarized in Table 1. Although the proposed PDA algorithm needs to store the two variables $\boldsymbol{w}_t$ and $\boldsymbol{s}_t$ as in RDA and ADAGRAD-RDA (see Section 3.2), the computational complexity is $\mathcal{O}(n)$ for $r = 1, 2$, which is a typical choice.

3.2. Relation to Prior Work

APFBS: One can apply the iterates $\boldsymbol{w}_t := \mathrm{prox}_{\eta\psi_t}^{\boldsymbol{Q}_t}(\boldsymbol{w}_{t-1} - \eta \boldsymbol{g}_t)$ to (1). This is actually a special case of APFBS [8, 9], which resembles the FOBOS algorithm [10] in the sense of using forward-backward splitting for online tasks. Note however that APFBS explicitly uses (the sum of multiple) squared-distance functions together with variable metrics, whereas FOBOS considers the ordinary least square cost $\varphi_t^{\mathrm{LS}}(\boldsymbol{w}) := \frac{1}{2}\left(y_t - \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_t\right)^2$ for regression with a fixed metric. APFBS is a projection-based forward-backward splitting algorithm, while PDA is based on RDA [11]. Figure 2 shows the difference between the forward-backward splitting method [29] and RDA. One can see that the effects of the proximity operator accumulate over the iteration. This actually increases the estimation biases, and APFBS therefore has a tradeoff between the strength of regularization and the estimation accuracy. RDA is free from the accumulation issue, yielding high estimation accuracy together with a high level of sparsity.

ADAGRAD-RDA: ADAGRAD [30] is one of the celebrated online learning methods in machine learning. The idea is to reduce the variance of the (sub)gradient vector by summing up the outer-products of the history of the (sub)gradient vectors to build a metric. The ADAGRAD algorithm was applied to two types of algorithms: RDA and the composite mirror descent [31, 32] (which is a generalization of FOBOS [10]). ADAGRAD-RDA has some similarities to the proposed method in the sense that both methods are based on RDA and employs variable metrics. The remarkable differences are, however, that the proposed method (i) seeks to minimize the "normalized" squared errors coming from the squared-distance cost (2) and (ii) utilizes a sparsity-promoting metric $\boldsymbol{Q}_t$. ADAGRAD-RDA uses the ordinary least square cost $\varphi_t^{\mathrm{LS}}(\boldsymbol{w})$. The gradient of $\varphi_t^{\mathrm{LS}}(\boldsymbol{w})$ can be disturbed by large inputs, which makes the algorithm unstable. Figure 3 shows the difference among the anti-gradient vectors $-\nabla\varphi_t^{\mathrm{LS}}(\boldsymbol{w})$, $-\nabla\varphi_t(\boldsymbol{w})$, and $-\nabla_{\boldsymbol{Q}_t}\varphi_t(\boldsymbol{w})$. The squared-distance cost (2) robustifies the gradient against large inputs. In addition, the metric $\boldsymbol{Q}_t$ guides the update direction towards the optimal point $\boldsymbol{w}_*$, leading to convergence acceleration.

4. NUMERICAL EXAMPLES

We show the efficacy of the proposed algorithm first in a simple sparse-system estimation problem and then in an acoustic echo cancellation problem. We compare the proposed algorithm with APA, PAPA, APFBS, RDA, and ADAGRAD-RDA. The RDA, ADAGRAD-RDA algorithms use $\varphi_t^{\mathrm{LS}}(\boldsymbol{w})$ and $\psi_t(\boldsymbol{w}) := \lambda \left|\left|\boldsymbol{w}\right|\right|_{\boldsymbol{I}_n,1}$. The other algorithms use $\varphi_t(\boldsymbol{w})$ in (2) and the weighted ℓ_1 norm in (13). In both experiments, the system mismatch $||\boldsymbol{w}_* - \boldsymbol{w}_t||_{\boldsymbol{I}_n}^2 / ||\boldsymbol{w}_*||_{\boldsymbol{I}_n}^2$ is used as a performance measure. Although there are some possible choice for $\boldsymbol{Q}_t$ such as in [3, 4, 23, 28], the metric in (12) are used for PAPA, APFBS, and PDA for fairness. In both experiments, the system mismatch is averaged over 300 independent trials.

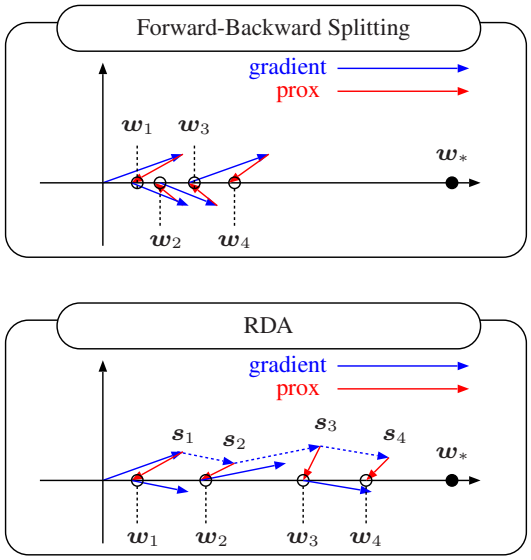


Fig. 2. Illustrations of forward-backward splitting and RDA.

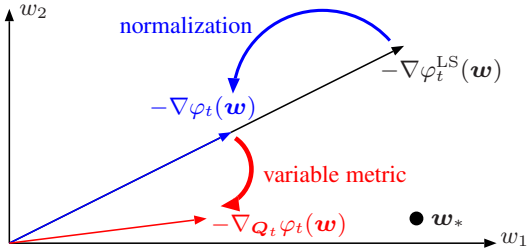


Fig. 3. The anti-gradients for a large input vector.

4.1. Sparse-System Estimation

We let the proportion of the zero components of the true coefficient vector $\boldsymbol{w}_* \in \mathbb{R}^{1000}$ be 90%, and the nonzero components are selected randomly from $[-4, 4]$. The noise ν_t is zero-mean i.i.d. Gaussian with variance 0.01. The input vector $\boldsymbol{x}_t \in \mathbb{R}^{1000}$ is randomly drawn from the i.i.d. uniform distribution over $[-2, 2]$. The parameters for each algorithm are chosen so that the speeds of initial convergence are nearly the same, and are shown in Table 2.

Figure 4(a) depicts the learning curves. One can see that the entire performance of PDA outperforms the other algorithms. The proportion of the zero components of the estimated coefficient vector is given as follows: APA (0%), PAPA (0%), APFBS (0%), RDA (11.6%), ADAGRAD-RDA (89%), and PDA (90%). PDA

Table 2. Parameters for sparse-system estimation.

Algorithms	η	λ	α	r	δ	ϵ
APA	0.16	-	-	1	10^{-5}	-
PAPA	0.14	-	0.8	1	10^{-5}	10^{-5}
APFBS	0.14	10^{-3}	0.8	1	10^{-5}	10^{-5}
RDA	0.01	10^{-3}	-	-	-	-
ADAGRAD	0.17	10^{-3}	-	-	-	-
PDA	0.13	3×10^{3}	0.8	1	10^{-5}	10^{-5}

Table 3. Parameters for echo cancellation.

Algorithms	η	λ	α	r	δ	ϵ
APA	0.3	-	-	2	10^{-15}	-
PAPA	0.3	-	0.2	2	10^{-15}	10^{-15}
APFBS	0.2	10^{-2}	0.3	2	10^{-15}	10^{-15}
RDA	1	10^{-4}	-	-	-	-
ADAGRAD	0.3	10^{-4}	-	-	-	-
PDA	0.2	25.5	0.3	2	10^{-15}	10^{-15}

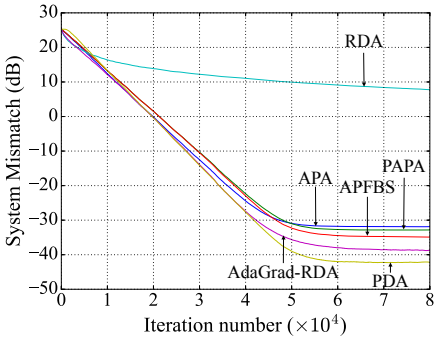
and ADAGRAD-RDA estimates the zero components accurately (see Section 3.2). Figure 4(b) depicts the proportion of zero components. One can see that PDA achieves an accurate sparsity-level remarkably faster than the other algorithms.

4.2. Echo Cancellation

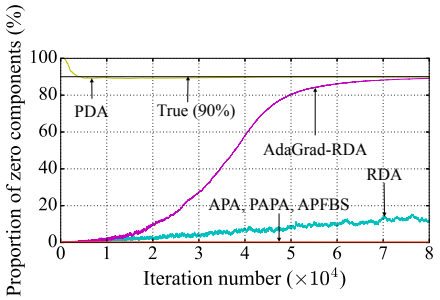
Figure 5(a) shows the amplitude of speech signal and Figure 5(b) shows the echo path used in the experiments. The sampling frequency of speech signal and echo path is 8000 Hz. The learning is stopped whenever the amplitude of input signal is below 10^{-4}. The parameters for each algorithm are shown in Table 3. The noise is zero-mean i.i.d. Gaussian with the signal noise ratio (SNR) 20 dB. Figure 6 shows the learning curves. The proportion of the zero components of the estimated coefficient vector is given as follows: APA (0%), PAPA (0%), APFBS (4.1%), RDA (51.4%), ADAGRAD-RDA (90.0%), and PDA (64.8%). Note here that the regularization parameter for each algorithm is chosen to give the best convergence behaviors. In this experiment, the mild sparsity of PDA yields a reasonably good convergence behavior. The use of the metric $\boldsymbol{Q}_t$ allows PAPA, APFBS, and PDA to attain fast initial convergence. In addition, PDA achieves the lowest system mismatch due to the strong regularization.

5. CONCLUSION

We proposed the projection-based dual averaging (PDA) algorithm, which features the input-vector normalization and the sparsity-seeking variable-metric. The input-vector normalization actually came from the squared-distance function to a closed convex set. Although the squared-distance function has been used in many adaptive filtering algorithms including NLMS, APA, APSM, and APFBS, its application to the dual averaging method has not been studied previously to the best of authors' knowledge. The similarities and dissimilarities between PDA and ADAGRAD-RDA were clarified. An application of PDA to an online regression problem was presented. The numerical examples demonstrated the better sparsity-seeking and learning properties for sparse-system estimation and the faster convergence for echo cancellation compared to the existing methods including ADAGRAD and APFBS. Our future works of particular interests include applications of PDA to machine learning tasks.

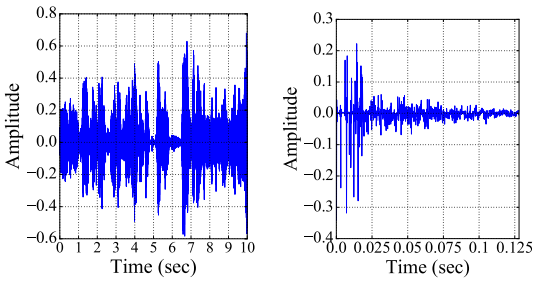


(a)



(b)

Fig. 4. Results for sparse-system estimation.



(a) Speech signal (b) Echo path

Fig. 5. Amplitudes of speech signal and echo path.

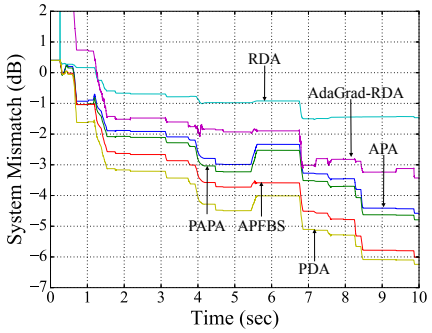


Fig. 6. System mismatch for echo cancellation.

6. REFERENCES

[1] H. Kushner and G. G. Yin, *Stochastic approximation and recursive algorithms and applications*. Springer Science & Business Media, 2003, vol. 35.

[2] K. Slavakis, S.-J. Kim, G. Mateos, and G. B. Giannakis, "Stochastic approximation vis-a-vis online learning for big data analytics [lecture notes]," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 124–129, 2014.

[3] D. L. Duttweiler, "Proportionate normalized least-mean-squares adaptation in echo cancelers," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 5, pp. 508–518, 2000.

[4] S. L. Gay, "An efficient, fast converging adaptive filter for network echo cancellation," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, vol. 1, 1998, pp. 394–398.

[5] S. Makino, Y. Kaneda, and N. Koizumi, "Exponentially weighted stepsize nlms adaptive filter based on the statistics of a room impulse response," *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 1, pp. 101–108, 1993.

[6] M. Yukawa, K. Slavakis, and I. Yamada, "Adaptive parallel quadratic-metric projection algorithms," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1665–1680, 2007.

[7] M. Yukawa and I. Yamada, "A unified view of adaptive variable-metric projection algorithms," *EURASIP J. Advances in Signal Processing*, vol. 2009, Article ID 589260, 13 pages, 2009.

[8] Y. Murakami, M. Yamagishi, M. Yukawa, and I. Yamada, "A sparse adaptive filtering using time-varying soft-thresholding techniques," in *Proc. IEEE ICASSP*, 2010, pp. 3734–3737.

[9] M. Yamagishi, M. Yukawa, and I. Yamada, "Acceleration of adaptive proximal forward-backward splitting method and its application to sparse system identification," in *Proc. IEEE ICASSP*, 2011, pp. 4296–4299.

[10] Y. Singer and J. C. Duchi, "Efficient learning using forward-backward splitting," in *Advances in Neural Information Processing Systems*, 2009, pp. 495–503.

[11] L. Xiao, "Dual averaging method for regularized stochastic learning and online optimization," in *Advances in Neural Information Processing Systems*, 2009, pp. 2116–2124.

[12] Y. Nesterov, "Primal-dual subgradient methods for convex problems," *Mathematical Programming*, vol. 120, no. 1, pp. 221–259, 2009.

[13] A. Kalai and S. Vempala, "Efficient algorithms for online decision problems," *Journal of Computer and System Sciences*, vol. 71, no. 3, pp. 291–307, 2005.

[14] J.-I. Nagumo and A. Noda, "A learning method for system identification," *IEEE Trans. Automatic Control*, vol. 12, no. 3, pp. 282–287, 1967.

[15] A. E. Albert and L. S. Gardner Jr., *Stochastic Approximation and Nonlinear Regression*. Cambridge MA: MIT Press, 1967.

[16] B. Widrow and M. E. Hoff, "Adaptive switching circuits," in *IRE WESCON convention record*, vol. 4, no. 1. New York, 1960, pp. 96–104.

[17] B. Widrow and S. D. Stearns, "Adaptive signal processing," *Englewood Cliffs, NJ, Prentice-Hall, Inc., 1985.*, vol. 1, 1985.

[18] I. Yamada and N. Ogura, "Adaptive projected subgradient method for asymptotic minimization of sequence of nonnegative convex functions," *Numer. Funct. Anal. Optim.*, vol. 25, no. 7&8, pp. 593–617, 2004.

[19] K. Slavakis, I. Yamada, and N. Ogura, "The adaptive projected subgradient method over the fixed point set of strongly attracting nonexpansive mappings," *Numerical Functional Analysis and Optimization*, vol. 27, no. 7-8, pp. 905–930, 2006.

[20] M. Yukawa, K. Slavakis, and I. Yamada, "Multi-domain adaptive learning based on feasibility splitting and adaptive projected subgradient method," *IEICE Trans. fundamentals of electronics, communications and computer sciences*, vol. 93, no. 2, pp. 456–466, 2010.

[21] T. Gansler, S. L. Gay, M. M. Sondhi, and J. Benesty, "Double-talk robust fast converging algorithms for network echo cancellation," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 6, pp. 656–663, 2000.

[22] J. Benesty, Y. A. Huang, J. Chen, and P. A. Naylor, "Adaptive algorithms for the identification of sparse impulse responses," *Selected Methods for Acoustic Echo and Noise Control*, vol. 5, pp. 125–153, 2006.

[23] C. Paleologu, J. Benesty, and S. Ciochin, "An improved proportionate NLMS algorithm based on the ℓ_0 norm," in *Proc. IEEE ICASSP*, 2010, pp. 309–312.

[24] K. Ozeki and T. Umeda, "An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties," *Electronics and Communications in Japan (Part I: Communications)*, vol. 67, no. 5, pp. 19–27, 1984.

[25] T. Hinamoto and S. Maekawa, "Extended theory of learning identification," *Electrical Engineering in Japan*, vol. 95, no. 5, pp. 101–107, 1975.

[26] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 1st ed. New York: NY: Springer, 2011.

[27] I. Yamada, M. Yukawa, and M. Yamagishi, "Minimizing the Moreau envelope of nonsmooth convex functions over the fixed point set of certain quasi-nonexpansive mappings," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, ser. Optimization and Its Applications, vol. 49. New York: Springer, 2011, pp. 345–390.

[28] M. Yukawa and I. Yamada, "Two product-space formulations for unifying multiple metrics in set-theoretic adaptive filtering," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, 2010, pp. 1010–1014.

[29] P.-L. Lions and B. Mercier, "Splitting algorithms for the sum of two nonlinear operators," *SIAM Journal on Numerical Analysis*, vol. 16, no. 6, pp. 964–979, 1979.

[30] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.

[31] P. Tseng, "On accelerated proximal gradient methods for convex-concave optimization. manuscript submitted to," *SIAM Journal on Optimization*, 2008.

[32] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari, "Composite objective mirror descent." in *Proc. COLT*, 2010, pp. 14–26.