Projection-based Dual Averaging For Stochastic Sparse Optimization

Asahi Ushio, Masahiro Yukawa

Dept. Electronics and Electrical Engineering, Keio University, Japan.
E-mail: ushio@ykw.elec.keio.ac.jp, yukawa@elec.keio.ac.jp

The 42nd ICASSP, @New Orleans, LA, USA
March 5-9, Mar, 2017

Introduction

Background Now, we have many kinds of data (SNS, IoT, Digital News) [1].
$\rightarrow$ **Machine learning and signal processing become more important !**

Challenges (demands for algorithms)

- Process **real-time streaming data**.
- Achieve **low complexity**.
- Deal with **high-dimensionality** and **sparsity** of data.

[1] McKinsey, "Big Data: The next frontier for innovation, competition, and productivity," 2011.

In Signal Processing ...

- **Squared distance cost:** Projection-based method ex) NLMS, APA, APFBS [2]
- **Change Geometry:** PAPA, Variable Metric [3]

In Machine Learning ...

- **Regularized Dual Averaging (RDA) type:** RDA [4]
- **Forward Backward Splitting (FBS) type:** FOBOS [5]
- **Change Geometry:** AdaGrad [6]

[2] Murakami *et al.*, 2010, [3] Yukawa *et al.*, 2009, [4] Xiao, 2009, [5] Singer *et al.*, 2009, [6] Duchi *et al.*, 2011

In Signal Processing ...

Background Now, we have many kinds of data (SNS, IoT, Digital News) [1].
$\rightarrow$ **Machine learning and signal processing become more important !**

Challenges (demands for algorithms)

- Process **real-time streaming data**.
- Achieve **low complexity**.
- Deal with **high-dimensionality** and **sparsity** of data.

[1] McKinsey, "Big Data: The next frontier for innovation, competition, and productivity," 2011.

In Signal Processing ...

- **Squared distance cost:** Projection-based method ex) NLMS, APA, APFBS [2]
- **Change Geometry:** PAPA, Variable Metric [3]

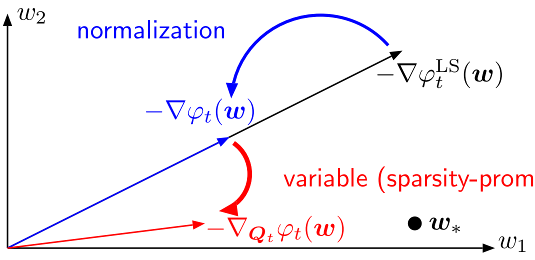
In Machine Learning ...

- **Regularized Dual Averaging (RDA) type:** RDA [4]
- **Forward Backward Splitting (FBS) type:** FOBOS [5]
- **Change Geometry:** AdaGrad [6]

[2] Murakami *et al.*, 2010, [3] Yukawa *et al.*, 2009, [4] Xiao, 2009, [5] Singer *et al.*, 2009, [6] Duchi *et al.*, 2011

An Illustration of Projection-based Method

in case of online regression



Ordinary least square cost.

$$\varphi_t^{\mathrm{LS}}(\boldsymbol{w}) := \frac{1}{2}\left(y_t - \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_t\right)^2$$

Squared distance cost with Q_t-metric.

$$\varphi_t(\boldsymbol{w}) = \frac{1}{2}\left(\frac{y_t - \langle \boldsymbol{w}, \boldsymbol{x}_t\rangle_{\boldsymbol{Q}_t}}{||\boldsymbol{x}_t||_{\boldsymbol{Q}_t}}\right)^2$$

$\boldsymbol{x}_t \in \mathbb{R}^n$: the input vector
$y_t \in \mathbb{R}$: the output
$\boldsymbol{w}_* \in \mathbb{R}^n$: the unknown vector

$\boldsymbol{Q}_t \in \mathbb{R}^{n\times n}$: a positive definite matrix
$\boldsymbol{Q}_t$-norm : $||\boldsymbol{w}||_{\boldsymbol{Q}_t} := \sqrt{\langle \boldsymbol{w}, \boldsymbol{w}\rangle_{\boldsymbol{Q}_t}}$,
$\langle \boldsymbol{w}, \boldsymbol{z}\rangle_{\boldsymbol{Q}_t} := \sqrt{\boldsymbol{w}^{\mathsf{T}}\boldsymbol{Q}_t\boldsymbol{z}}$ for $\boldsymbol{w}, \boldsymbol{z} \in \mathbb{R}^n$

In Machine Learning ...

Background Now, we have many kinds of data (SNS, IoT, Digital News) [1].
$\rightarrow$ **Machine learning and signal processing become more important !**

Challenges (demands for algorithms)

- Process **real-time streaming data**.
- Achieve **low complexity**.
- Deal with **high-dimensionality** and **sparsity** of data.

[1] McKinsey, "Big Data: The next frontier for innovation, competition, and productivity," 2011.

In Signal Processing ...

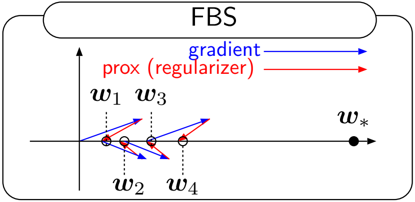
- **Squared distance cost:** Projection-based method ex) NLMS, APA, APFBS [2]
- **Change Geometry:** PAPA, Variable Metric [3]

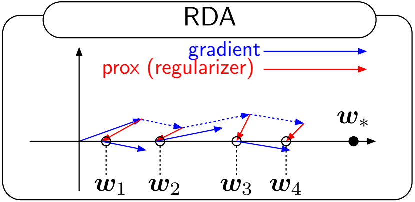
In Machine Learning ...

- **Regularized Dual Averaging (RDA) type:** RDA [4]
- **Forward Backward Splitting (FBS) type:** FOBOS [5]
- **Change Geometry:** AdaGrad [6]

[2] Murakami *et al.*, 2010, [3] Yukawa *et al.*, 2009, [4] Xiao, 2009, [5] Singer *et al.*, 2009, [6] Duchi *et al.*, 2011

FBS type versus RDA type





- **FBS:** The effects of the proximity operator accumulate over the iterations.

Tradeoff between the sparsity level and the estimation accuracy.

- **RDA:** FREE from the accumulation.

A high level of sparsity comes with high estimation accuracy !

Abstract of This Work

Motivation

1. **RDA:** Sparse solution.
2. **Squared distance cost:** Stable adaptation.
3. **Variable metric:** Promoting sparsity to improve the performance.

Combination of 3 properties.

Sparsity-promoting and stable learning !

Proposed Algorithm

Projection-based Dual Averaging (PDA)

- Features: **RDA** with **Squared distance cost**, employing **variable metric**.
- Show efficacy by numerical examples (simulated data, real data).

Preliminaries

Problem Setting : an online regularized optimization

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

ψ_t,φ_t : a possibly nonsmooth function
$\boldsymbol{w}$: supposed to be sparse or compressible

Basic Stochastic Optimization Methods

- **SGD**

$$\boldsymbol{w}_t := \boldsymbol{w}_{t-1} - \eta\nabla\varphi_t(\boldsymbol{w}_{t-1}), \quad \eta > 0 \tag{2}$$

- **Dual Averaging** [1]

$$\boldsymbol{w}_t := \arg\min_{\boldsymbol{w}\in\mathbb{R}^n} \left(\left\langle \frac{\sum_{i=1}^t \nabla\varphi_{i-1}(\boldsymbol{w}_{i-1})}{t}, \boldsymbol{w} \right\rangle + \mu_t h(\boldsymbol{w}) \right), \mu_t = \mathcal{O}\left(\frac{1}{\sqrt{t}}\right) \tag{3}$$

$h(\boldsymbol{w})$: the prox-function

[1] Y. Nesterov, "Primal-dual subgradient methods for convex problems", Mathematical Programming, 2009.

Projection-based Methods

Cost Function of Projection-based Methods

- **The Q_t-metric distance cost (normalized MSE in regression case).**

$$\varphi_t(\boldsymbol{w}) := \frac{1}{2} d^2_{\boldsymbol{Q}_t}(\boldsymbol{w}, C_t) \tag{4}$$

$$d_{\boldsymbol{Q}_t}(\boldsymbol{w}, C_t) := \min_{\boldsymbol{z} \in C_t} ||\boldsymbol{w} - \boldsymbol{z}||_{\boldsymbol{Q}_t} \tag{5}$$

$\boldsymbol{Q}_t \in \mathbb{R}^{n \times n}$: a positive definite matrix

$C_t \subset \mathbb{R}^n$: a closed convex set

$\boldsymbol{Q}_t$-norm for $\boldsymbol{w}, \boldsymbol{z} \in \mathbb{R}^n$: $||\boldsymbol{w}||_{\boldsymbol{Q}_t} := \sqrt{\langle \boldsymbol{w}, \boldsymbol{w} \rangle_{\boldsymbol{Q}_t}}, \langle \boldsymbol{w}, \boldsymbol{z} \rangle_{\boldsymbol{Q}_t} := \sqrt{\boldsymbol{w}^\mathsf{T} \boldsymbol{Q}_t \boldsymbol{z}}$

Gradient Calculation

- **The Q_t-gradient of φ_t**

$$\boldsymbol{g}_t := \nabla_{\boldsymbol{Q}_t} \varphi_t(\boldsymbol{w}_{t-1}) = \boldsymbol{w}_{t-1} - P^{\boldsymbol{Q}_t}_{C_t}(\boldsymbol{w}_{t-1}), \quad \boldsymbol{w}_{t-1} \in \mathbb{R}^n. \tag{6}$$

- **The Q_t-projection onto C_t**

$$P^{\boldsymbol{Q}_t}_{C_t}(\boldsymbol{w}) := \arg\min_{\boldsymbol{z} \in \mathcal{C}_t} ||\boldsymbol{w} - \boldsymbol{z}||_{\boldsymbol{Q}_t}. \tag{7}$$

Proposed Algorithm

Projection-based Dual Averaging (PDA)

$$
\begin{aligned}
\boldsymbol{w}_t &:= \underset{\boldsymbol{w}\in\mathbb{R}^n}{\arg\min}\left(\langle \boldsymbol{s}_t, \boldsymbol{w}\rangle_{\boldsymbol{Q}_t} + \frac{\|\boldsymbol{w}\|^2_{\boldsymbol{Q}_t}}{2\eta} + \psi_t(\boldsymbol{w})\right)\\
&= \underset{\boldsymbol{w}\in\mathbb{R}^n}{\arg\min}\left(\eta\psi_t(\boldsymbol{w}) + \frac{1}{2}\|\boldsymbol{w}+\eta\boldsymbol{s}_t\|^2_{\boldsymbol{Q}_t}\right)\\
&= \mathrm{prox}^{\boldsymbol{Q}_t}_{\eta\psi_t}(-\eta\boldsymbol{s}_t), \quad \boldsymbol{s}_t = \sum_{i=1}^{t}\nabla_{\boldsymbol{Q}_t}\varphi_{i-1}(\boldsymbol{w}_{i-1}), \quad \eta\in[0,2]
\end{aligned} \tag{8}
$$

- **The proximity operator**

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

Application to Online Regression

Problem Setting : Online Regression

$$y_t := \boldsymbol{w}_*^\top \boldsymbol{x}_t + \nu_t \in \mathbb{R}$$

$\boldsymbol{x}_t \in \mathbb{R}^n$: the input vector
$y_t \in \mathbb{R}$: the output
$\boldsymbol{w}_* \in \mathbb{R}^n$: the unknown vector
$\nu_t \in \mathbb{R}$: the additive white noise

Definition of the Projection

- **The linear variety**

$$C_t := \arg\min_{\boldsymbol{w}\in\mathbb{R}^n} \left|\left|\boldsymbol{X}^\top \boldsymbol{w} - \boldsymbol{y}_t\right|\right|_{\boldsymbol{I}_n}. \tag{10}$$

$\boldsymbol{X}_t := [\boldsymbol{x}_t \cdots \boldsymbol{x}_{t-r+1}] \in \mathbb{R}^{n\times r}$
$\boldsymbol{y}_t := [y_t, \ldots, y_{t-r+1}]^\top \in \mathbb{R}^r$

- **The projection onto C_t**

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

- **The Moore-Penrose pseudo-inverse $\boldsymbol{X}_t^\dagger$**

$$\boldsymbol{X}_t(\boldsymbol{X}_t^\top \boldsymbol{Q}_t^{-1}\boldsymbol{X}_t + \delta \boldsymbol{I}_n)^{-1}, \quad \delta > 0.$$

Design of Metric $\boldsymbol{Q}_t$ & Regularizer ψ_t

Design of metric $\boldsymbol{Q}_t$ (following [Yukawa *et al.* 2010])

$$\boldsymbol{Q}_t := \frac{\alpha}{n}\boldsymbol{I}_n + \frac{1-\alpha}{S_t}\tilde{\boldsymbol{Q}}_t^{-1}, \quad \alpha \in [0,1]. \tag{12}$$

- $\tilde{\boldsymbol{Q}}_t := \mathrm{diag}(|w_{t-1,1}|, \ldots, |w_{t-1,n}|) + \epsilon\boldsymbol{I}_n$ for some $\epsilon > 0$
- $S_t := \sum_{i=1}^{n}(|w_{t-1,i}| + \epsilon)^{-1}$ to reduce $\mathrm{tr}(\boldsymbol{I}_n/n) = \mathrm{tr}(\tilde{\boldsymbol{Q}}_t^{-1}/S_t) = 1$

The Regularization Term ψ_t

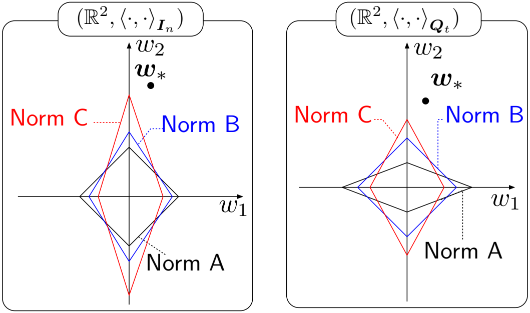
$$\psi_t(\boldsymbol{w}) = \lambda\,||\boldsymbol{w}||_{\boldsymbol{Q}_t^2,1} := \lambda\sum_{i=1}^{n} q_{t,i}^2|w_i|, \quad \lambda > 0. \tag{13}$$

- $\boldsymbol{Q}_t := \mathrm{diag}(q_{t,1} \ldots, q_{t,n}) \in \mathbb{R}^{n\times n}$
- **The proximity operator** for the ψ_t in (13)

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

- $\{\boldsymbol{e}_i\}_{i=1}^n$: the standard basis of $\mathbb{R}^n$

The Hilbert spaces $(\mathbb{R}^2, \langle\cdot,\cdot\rangle_{\boldsymbol{I}_n})$ and $(\mathbb{R}^2, \langle\cdot,\cdot\rangle_{\boldsymbol{Q}_t})$



- **Norm A**
  $||\boldsymbol{w}||_{\boldsymbol{I}_n,1} := \sum_{i=1}^n |w_i|$
- **Norm B**
  $||\boldsymbol{w}||_{\boldsymbol{Q}_t^1,1} := \sum_{i=1}^n q_{t,i}|w_i|$
- **Norm C (we employ)**
  $||\boldsymbol{w}||_{\boldsymbol{Q}_t^2,1} := \sum_{i=1}^n q_{t,i}^2|w_i|$

- **Norm A** gives a **fat** unit ball in $(\mathbb{R}^2, \langle\cdot,\cdot\rangle_{\boldsymbol{Q}_t})$:
  The proximity operator shrinks the large component more than the small one.

Undesirable bias.

- **Norm C** gives a **tall** unit ball in $(\mathbb{R}^2, \langle\cdot,\cdot\rangle_{\boldsymbol{Q}_t})$:
  Shrink the small components more.

Reduce the bias !

Relation to Prior Work

SGD type: SGD, **NLMS**, **APA**, **PAPA**

Sparsity-promoting: FBS (Forward Backward Splitting) type

FOBOS [1], AdaGrad-FBS [2], **APFBS** [3]

Dual Averaging type: Dual Averaging [4]

Sparsity-promoting: RDA (Regularized Dual Averaging) type

RDA [5], AdaGrad-RDA [2]

*bold : projection-based method

[1] Singer *et al.*, 2009 [2] Duchi *et al.*, 2011 [3] Murakami *et al.*, 2010 [4] Nesterov, 2009 [5] Xiao, 2009

	Ordinal Cost Function	Projection-based
FBS type	FOBOS, AdaGrad-FBS	**APFBS**
RDA type	RDA, AdaGrad-RDA	

Relation to Prior Work

SGD type: SGD, **NLMS**, **APA**, **PAPA**

Sparsity-promoting: FBS (Forward Backward Splitting) type

FOBOS [1], AdaGrad-FBS [2], **APFBS** [3]

Dual Averaging type: Dual Averaging [4]

Sparsity-promoting: RDA (Regularized Dual Averaging) type

RDA [5], AdaGrad-RDA [2], **PDA**

*bold : projection-based method

[1] Singer *et al.*, 2009 [2] Duchi *et al.*, 2011 [3] Murakami *et al.*, 2010 [4] Nesterov, 2009 [5] Xiao, 2009

	Ordinal Cost Function	Projection-based
FBS type	FOBOS, AdaGrad-FBS	**APFBS**
RDA type	RDA, AdaGrad-RDA	**PDA**

Numerical Example

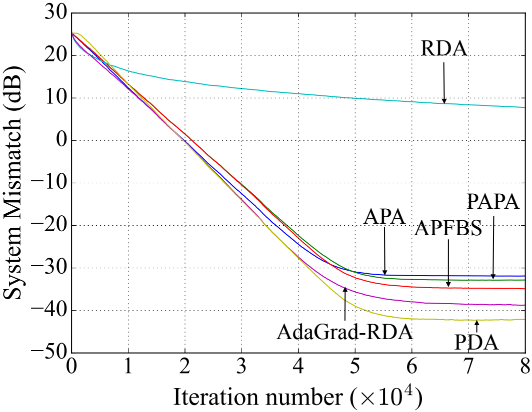
- **Experiment:**
  1. Sparse-System Estimation (Simulated Data)

     Model $y_t := \boldsymbol{w}_*^{\mathsf{T}} \boldsymbol{x}_t + \nu_t \in \mathbb{R}$
     - The proportion of the zero components of $\boldsymbol{w}_* \in \mathbb{R}^{1000}$ is 90%.
     - The noise ν_t is zero-mean i.i.d. Gaussian with variance 0.01.
  2. Echo Cancellation (Real Data)
     - The sampling frequency of speech signal and echo path is 8000 Hz.
     - The learning is stopped whenever the amplitude of input is below 10^{-4}.
     - The noise is zero-mean i.i.d. Gaussian with the signal noise ratio 20 dB.
- **Compared Algorithms:** APA, PAPA, APFBS, RDA, AdaGrad-RDA

	APA	PAPA	APFBS	RDA	AdaGrad-RDA	PDA
Cost	$\varphi_t(\boldsymbol{w})$	$\varphi_t(\boldsymbol{w})$	$\varphi_t(\boldsymbol{w})$	$\varphi_t^{\mathrm{LS}}(\boldsymbol{w})$	$\varphi_t^{\mathrm{LS}}(\boldsymbol{w})$	$\varphi_t(\boldsymbol{w})$
Reg	-	-	$\lambda \|\|\boldsymbol{w}\|\|_{\boldsymbol{Q}_t^2,1}$	$\lambda \|\|\boldsymbol{w}\|\|_{\boldsymbol{I}_n,1}$	$\lambda \|\|\boldsymbol{w}\|\|_{\boldsymbol{I}_n,1}$	$\lambda \|\|\boldsymbol{w}\|\|_{\boldsymbol{Q}_t^2,1}$

Experiment 1: Sparse-System Estimation

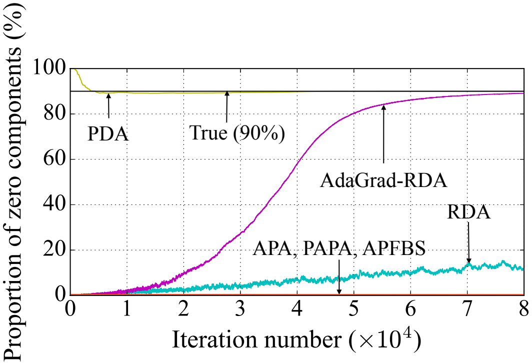
System Mismatch



- $||\boldsymbol{w}_* - \boldsymbol{w}_t||^2_{\boldsymbol{I}_n} / ||\boldsymbol{w}_*||^2_{\boldsymbol{I}_n}$, $\boldsymbol{w}_*$ is true parameter, $\boldsymbol{w}_t$ is estimation at t.
- **PDA shows the best performance.**

Experiment 1: Sparse-System Estimation

Proportion of the Zero Components of the Estimated Coefficient Vector

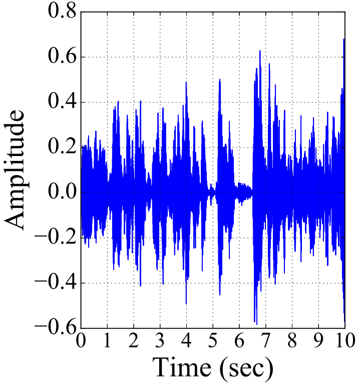


Algorithms	APA	PAPA	APFBS	RDA	AdaGrad-RDA	PDA
Proportion	0%	0%	0%	11.6%	89%	90%

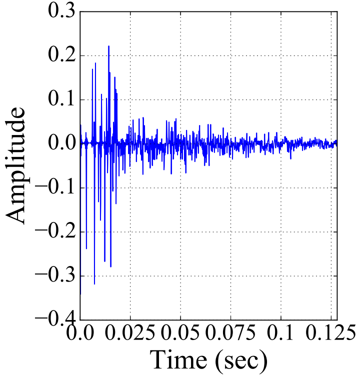
- **PDA achieve accurate sparsity.**

Experiment 2: Echo Cancellation

Amplitudes of Speech Signal and Echo Path



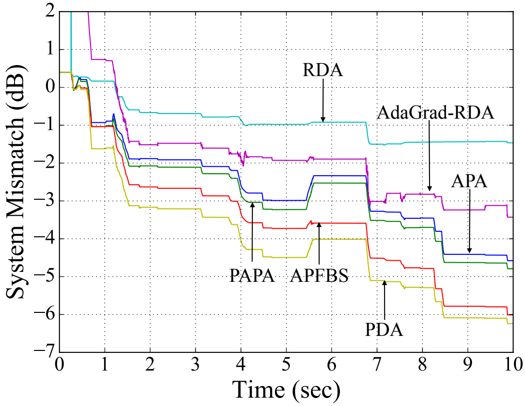
(a) Speech signal



(b) Echo path

Experiment 2: Echo Cancellation

System Mismatch



- **PDA shows the best performance.**

Conclusion

Conclusion

- We proposed the **projection-based dual averaging (PDA)** algorithm.
  - **projection-based:** Input-vector normalization and the sparsity-seeking variable-metric
  - **RDA:** Better sparsity-seeking.
- An application of PDA to an **online regression** problem was presented.
  - The numerical examples demonstrated the **better sparsity-seeking** and **learning properties**.

Future Work

- Application to machine learning problems (classification).
- Self-tuning method for λ and α.

Parameters for Experiments

Table: 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: 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}