A Fixed-Point Analysis of Regularized Dual Averaging Under Static Scenarios

Masahiro Yukawa*† Isao Yamada‡
* Department of Electronics and Electrical Engineering, Keio University, JAPAN
† Center for Advanced Intelligence Project, RIKEN, JAPAN
‡ Department of Information and Communications Engineering, Tokyo Institute of Technology, JAPAN

***Abstract*—In this paper, we analyze the properties of a fixed point of a certain mapping that is implicitly used in each of the regularized dual averaging (RDA) and projection-based RDA (PDA) algorithms. It turns out that, if the loss function has a nonexpansive (1-Lipschitz) gradient such as in the case of a half squared-distance function, RDA converges to a minimizer of the penalized loss function under a restrictive condition. Meanwhile, the fixed point for PDA gives a minimizer of the 'unpenalized' loss function. Some simulation studies are also presented to support the theoretical findings.**

I. Introduction

The regularized dual averaging (RDA) algorithm [1] and the adaptive proximal forward-backward splitting (APFBS) algorithm [2] (or FOBOS [3]) are two major lines of research on regularized stochastic optimization algorithms. APFBS, or FOBOS, is an adaptive/online extension of the proximal forward-backward splitting method (also known as the proximal gradient method), which is a particular case of the Krasnoselskii-Mann (KM) iterate and of which the convergence mechanism is thus transparent based on the fixed-point characterization of nonexpansive mapping (see [4] for instance). On the other hand, RDA is motivated by the dual averaging algorithm of Nesterov [5], and its convergence properties have been studied only in the stochastic sense. Motivated by the success of the projection-based methods for adaptive filtering [6–9], the projection-based RDA (PDA) algorithm has been proposed [10, 11], employing a half squared-distance loss together with a variable-metric. It has been shown that, when applied to sparse system identification, PDA exhibits better convergence behaviours as well as a better sparsity-seeking property. To understand the basic principle of RDA/PDA, it is of great interest to study how those algorithms can be seen from the fixed-point theoretic viewpoint in the static scenario.

In this paper, we analyze the properties of a fixed point of a certain mapping that is implicitly used in each of RDA and PDA. It turns out that, if the loss function has a nonexpansive (i.e., 1-Lipschitz) gradient such as in the case of the half squared-distance function, RDA converges to a minimizer of the penalized loss function under a restrictive condition. Meanwhile, the fixed point for PDA gives a minimizer of the 'unpenalized' loss function, which is independent from the regularizer. Simulation results support the theoretical findings.

II. Preliminaries

A. Mathematical Tools

Let $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ be a real Hilbert space equipped with inner product $\langle\cdot,\cdot\rangle$. We denote its induced norm by $\|\cdot\|$. A convex function f satisfying $\mathrm{dom} f := \{x \in \mathcal{H} \mid f(x) < \infty\} \neq \emptyset$ is called a *proper convex* function.[1] A function $f : \mathcal{H} \to (-\infty, \infty]$ is said to be *lower semicontinuous* on $\mathcal{H}$ if the level set $\mathrm{lev}_{\leq a} f := \{x \in \mathcal{H} : f(x) \leq a\}$ is closed for every $a \in \mathbb{R}$. We denote by $I : \mathcal{H} \to \mathcal{H}$ the identity operator which maps any vector $\boldsymbol{x} \in \mathcal{H}$ to the $\boldsymbol{x}$ itself.

Definition 1 (Lipschitz continuity and nonexpansivity)**.** *A mapping $T : \mathcal{H} \to \mathcal{H}$ is called Lipschitz continuous with constant $\kappa > 0$ (or κ-Lipschitz for short) if for any $x, y \in \mathcal{H}$*

$$\|T(x) - T(y)\| \leq \kappa \|x - y\|. \tag{1}$$

A 1-Lipschitz mapping is specially called nonexpansive.

Lipschitz continuity implies continuity in the ordinary sense since $\|x - y\| \to 0$ clearly implies $\|T(x) - T(y)\| \to 0$ by definition.

Definition 2 (Fixed point)**.** *A point that is "fixed" under the operation of $T : \mathcal{H} \to \mathcal{H}$ (i.e. a point $x \in \mathcal{H}$ such that $T(x) = x$) is called a fixed point of T. We denote the set of all fixed points of T by $\mathrm{Fix}(T)$.*

Definition 3 (Averaged nonexpansivity)**.** *A mapping $T : \mathcal{H} \to \mathcal{H}$ is called α-averaged nonexpansive for a constant $\alpha \in (0, 1)$ if there exists a nonexpansive mapping $N : \mathcal{H} \to \mathcal{H}$ such that $T = (1 - \alpha)I + \alpha N$.*

Definition 4 (Proximity operator [4, 12])**.** *Given any proper lower-semicontinuous convex function $f : \mathcal{H} \to (-\infty, \infty]$, the proximity operator of f of index $\gamma > 0$ is defined as*

$$\mathrm{prox}_{\gamma f}(x) := \operatorname*{argmin}_{y \in \mathcal{H}} \left(f(y) + \frac{1}{2\gamma} \|x - y\|^2 \right), \quad x \in \mathcal{H}.$$

Definition 5 (Subdifferential [4, 13])**.** *Given $x \in \mathcal{H}$ and proper lower-semicontinuous convex function $f : \mathcal{H} \to (-\infty, \infty]$,*

$$\partial f(x) := \{z \in \mathcal{H} \mid \langle y - x, z\rangle + f(x) \leq f(y), \quad \forall y \in \mathcal{H}\} \tag{2}$$

[1] A subset $S \subset \mathcal{H}$ is said to be convex if $\alpha x + (1 - \alpha)y \in S$ for all $(x, y, \alpha) \in S \times S \times [0, 1]$. A function $f : \mathcal{H} \to (-\infty, \infty] := \mathbb{R} \cup \{\infty\}$ is said to be convex on $\mathcal{H}$ if $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$ for all $(x, y, \alpha) \in \mathrm{dom} f \times \mathrm{dom} f \times [0, 1]$, where $\mathrm{dom} f := \{x \in \mathcal{H} \mid f(x) < \infty\}$. The function f is called strictly convex if the inequality of convex function holds with strict inequality whenever $x \neq y$.

is called the subdifferential of f at x. If f is continuous, it is ensured that $\partial f(x) \neq \emptyset$.

Definition 6 (Indicator function). *Given a nonempty closed convex set $C \subset \mathcal{H}$, define the indicator function $\iota_C(x) := \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{if } x \notin C. \end{cases}$ The function ι_C is lower semicontinuous because $\mathrm{lev}_{\leq a}\iota_C = C$ if $a \geq 0$ and $\mathrm{lev}_{\leq a}\iota_C = \emptyset$ if $a < 0$, although it is clearly discontinuous at the boundary of C.*

Fact 1 (On proximity operator [4, 13]).
1) $\mathrm{prox}_{\gamma f} = (I + \gamma\partial f)^{-1}$ [13].
2) $\mathrm{prox}_{\iota_C} = P_C : \mathcal{H} \to C,\ x \mapsto \mathrm{argmin}_{y \in C} \|x-y\|$ *is the metric projection operator onto the closed convex set $C \neq \emptyset$.*
3) *The proximity operator is firmly nonexpansive; i.e., 1/2-averaged nonexpansive, with $\mathrm{Fix}(\mathrm{prox}_f) = \mathrm{argmin}_{x\in\mathcal{H}} f(x)$. In the case of metric projection, in particular, $\mathrm{Fix}(P_C) = \mathrm{argmin}_{x\in\mathcal{H}} \iota_C(x) = C$.*

Fact 2 (On nonexpansive mapping [4, 13]).
1) *T is nonexpansive if and only if $-T$ is nonexpansive.*
2) *Given any nonexpansive mappings $T_1 : \mathcal{H} \to \mathcal{H}$ and $T_2 : \mathcal{H} \to \mathcal{H}$, their composition $T_2 \circ T_1$ is also nonexpansive.*
3) *The following three statements are equivalent: (a) T is firmly nonexpansive, (b) $I - T$ is firmly nonexpansive, (c) $2T - I$ is nonexpansive.*

Theorem 1 (Special case of KM iterate [4, 13]). *Let $T : \mathcal{H} \to \mathcal{H}$ be a nonexpansive mapping with $\mathrm{Fix}(T) \neq \emptyset$. Also let $(\alpha_t)_{t\in\mathbb{N}}$ is a sequence in [0,1] such that $\sum_{t\in\mathbb{N}} \alpha_t(1-\alpha_t) = \infty$. Then, for any initial point $w_0 \in \mathcal{H}$, the sequence $(w_t)_{t\in\mathbb{N}}$ generated by*

$$w_{t+1} := (1-\alpha_t)w_t + \alpha_t T(w_t) \tag{3}$$

converges weakly to a point $w^* \in \mathrm{Fix}(T)$.[2]

B. Regularized Stochastic Optimization Problem

We consider the following regularized stochastic optimization problems:

$$\min_{\boldsymbol{w}\in\mathbb{R}^n} \mathbb{E}_z\left[f(\boldsymbol{w}, z)\right] + \psi(\boldsymbol{w}), \tag{4}$$

where the first term is the expectation of the convex loss function $f(\boldsymbol{w}, z)$ with respect to the pair $z := (\boldsymbol{x}, y) \in \mathbb{R}^n \times \mathbb{R}$ of input $\boldsymbol{x}$ and output y drawn from an unknown underlying distribution, and $\psi(\boldsymbol{w})$ is the proper convex regularizer which is assumed lower-semicontinuous. In practice, the following empirical loss at each time instant $t \in \mathbb{N}$ is commonly considered:

$$\min_{\boldsymbol{w}\in\mathbb{R}^n} \frac{1}{t}\sum_{\tau=1}^{t}\left[\varphi_\tau(\boldsymbol{w})\right] + \psi(\boldsymbol{w}), \tag{5}$$

where $\varphi_\tau(\boldsymbol{w}) := f(\boldsymbol{w}, z_\tau)$ is assumed differentiable with the observation $z_\tau := (\boldsymbol{x}_\tau, y_\tau) \in \mathbb{R}^n \times \mathbb{R}$ of z at time instant $\tau = 1, 2, \cdots, t$.[3] In this case, $\mathrm{dom}\varphi_\tau = \mathbb{R}^n$. The estimate of an optimal $\boldsymbol{w}$ at time τ is denoted by $\boldsymbol{w}_\tau := [w_{\tau,1}, w_{\tau,2}, \cdots, w_{\tau,n}]^\mathsf{T} \in \mathbb{R}^n$.

III. Convergence Analysis of RDA Algorithm under Static Scenario

A. RDA Algorithm for $\beta_t = t$

Define the sum of the history of the gradients as

$$\boldsymbol{s}_t := \sum_{\tau=1}^{t} \nabla\varphi_\tau(\boldsymbol{w}_{\tau-1}) = \boldsymbol{s}_{t-1} + \nabla\varphi_t(\boldsymbol{w}_{t-1}), \quad t \in \mathbb{N}, \tag{6}$$

with $\boldsymbol{s}_0 := \boldsymbol{0}$. Let $(\beta_t)_{t\in\mathbb{N}} \subset (0,\infty)$ be a nondecreasing sequence. Also let $h(\boldsymbol{w})$ be a strongly-convex continuous function (called a prox-function) satisfying $\mathrm{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} h(\boldsymbol{w}) \subset \mathrm{argmin}_{\boldsymbol{y}\in\mathbb{R}^n} \psi(\boldsymbol{y})$. The RDA algorithm is then given by [1]

$$\boldsymbol{w}_t := \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n}\left(\left\langle \frac{\boldsymbol{s}_t}{t}, \boldsymbol{w}\right\rangle + \frac{\beta_t}{t}h(\boldsymbol{w}) + \psi(\boldsymbol{w})\right). \tag{7}$$

In the present study, we consider the case of $\beta_t := t$ and $h(\boldsymbol{w}) := \|\boldsymbol{w}\|^2/2 = \frac{1}{2}\sum_{i=1}^n w_i^2$, which is a typical choice for $\psi(\boldsymbol{w}) := \|\boldsymbol{w}\|_1 := \sum_{i=1}^n |w_i|$. In this case, (7) reduces to

$$\begin{aligned} \boldsymbol{w}_t &= \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n}\left(\left\langle \frac{\boldsymbol{s}_t}{t}, \boldsymbol{w}\right\rangle + \frac{1}{2}\|\boldsymbol{w}\|^2 + \psi(\boldsymbol{w})\right) \\ &= \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n}\left(\frac{1}{2}\left\|\boldsymbol{w} + \frac{\boldsymbol{s}_t}{t}\right\|^2 + \psi(\boldsymbol{w})\right) \\ &= \mathrm{prox}_\psi\left(-\frac{\boldsymbol{s}_t}{t}\right). \end{aligned} \tag{8}$$

B. Convergence Analysis

To make the analysis tractable, we consider the static scenario in which the loss function φ_τ does not change in time. We thus drop the time index of the loss function and denote it by φ. Define the mapping

$$T_t := \left(1 - \frac{1}{t}\right)I + \frac{1}{t}(-\nabla\varphi \circ \mathrm{prox}_\psi). \tag{9}$$

Then, the following proposition holds.

Proposition 1. *The sequence $(\boldsymbol{w}_t)_{t\in\mathbb{N}}$ generated by*

$$\begin{aligned} \boldsymbol{w}_t &:= \mathrm{prox}_\psi(\boldsymbol{\zeta}_t) \\ \boldsymbol{\zeta}_t &:= T_t(\boldsymbol{\zeta}_{t-1}), \quad \boldsymbol{\zeta}_0 := \boldsymbol{0}, \end{aligned} \tag{10}$$

coincides with the one generated by (8)*, which is the RDA algorithm for $h(\boldsymbol{w}) := \|\boldsymbol{w}\|^2/2$ and $\beta_t := t$.*

Proof: One can verify that

$$\begin{aligned} \boldsymbol{\zeta}_t &= \left(1 - \frac{1}{t}\right)\boldsymbol{\zeta}_{t-1} - \frac{1}{t}\nabla\varphi(\mathrm{prox}_\psi \boldsymbol{\zeta}_{t-1}) \\ &= \frac{t-1}{t}\boldsymbol{\zeta}_{t-1} - \frac{1}{t}\nabla\varphi(\boldsymbol{w}_{t-1}) \\ &= -\frac{1}{t}\left(-(t-1)\boldsymbol{\zeta}_{t-1} + \nabla\varphi(\boldsymbol{w}_{t-1})\right) \\ &= -\frac{\boldsymbol{s}_t}{t}. \end{aligned} \tag{11}$$

■

[2] A sequence $(w_t)_{t\in\mathbb{N}}$ is said to be weakly convergent to $w^* \in \mathcal{H}$ if $\lim_{t\to\infty}\langle w_t - w^*, y\rangle = 0$ for any $y \in \mathcal{H}$. In the finite dimensional case, the weak convergence coincides with the strong convergence (i.e., $\lim_{t\to\infty}\|w_t - w^*\| = 0 \Leftrightarrow \lim_{t\to\infty}\langle w_t - w^*, y\rangle = 0$ for any $y \in \mathcal{H}$).

[3] Although a time-dependent regularizer is considered in [10, 11], we solely consider the fixed regularizer in the present study for the sake of tractability.

Now, the question is what is the characterization of the fixed point. The hope is that its associated point $\boldsymbol{w}^* = \mathrm{prox}_\psi(\boldsymbol{\zeta}^*)$ is a solution to the minimization problem in (5), which in the current static case is a minimizer of $\varphi(\boldsymbol{w}) + \psi(\boldsymbol{w})$. This however holds only in a restrictive condition, as clarified in the theorem below together with its following arguments.

Theorem 2 (Fixed point of T_t)**.** *The following statements hold.*

1) $\mathrm{Fix}(T_t) = \mathrm{Fix}(-\nabla\varphi \circ \mathrm{prox}_\psi)$.
2) *Assume that* $\mathrm{Fix}(-\nabla\varphi\circ\mathrm{prox}_\psi) \neq \emptyset$. *Then, given a fixed point* $\boldsymbol{\zeta}^* \in \mathrm{Fix}(-\nabla\varphi \circ \mathrm{prox}_\psi)$, *the following statements are equivalent.*
   a) $\mathrm{prox}_\psi \boldsymbol{\zeta}^* \in \mathrm{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \varphi(\boldsymbol{w}) + \psi(\boldsymbol{w})$.
   b) $\boldsymbol{\zeta}^* \in \partial\psi(\mathrm{prox}_\psi \boldsymbol{\zeta}^*)$.
   c) $\mathrm{prox}_\psi \left(\boldsymbol{\zeta}^* + \mathrm{prox}_\psi \boldsymbol{\zeta}^*\right) = \mathrm{prox}_\psi \boldsymbol{\zeta}^*$.

Proof: Item 1 can be verified by observing that

$$
\begin{aligned}
T_t(\boldsymbol{\zeta}^*) = \boldsymbol{\zeta}^* &\Leftrightarrow \boldsymbol{\zeta}^* - \frac{1}{t}(\boldsymbol{\zeta}^* + \nabla\varphi \circ \mathrm{prox}_\psi(\boldsymbol{\zeta}^*)) = \boldsymbol{\zeta}^* \\
&\Leftrightarrow -\nabla\varphi \circ \mathrm{prox}_\psi(\boldsymbol{\zeta}^*) = \boldsymbol{\zeta}^*. \qquad (12)
\end{aligned}
$$

Item 2 can be verified as follows:

$$
\begin{aligned}
&\mathrm{prox}_\psi \boldsymbol{\zeta}^* \in \underset{\boldsymbol{w}\in\mathbb{R}^n}{\mathrm{argmin}}\, \varphi(\boldsymbol{w}) + \psi(\boldsymbol{w}) \\
&\Leftrightarrow \mathbf{0} \in \partial(\varphi + \psi)(\mathrm{prox}_\psi \boldsymbol{\zeta}^*) = \nabla\varphi(\mathrm{prox}_\psi \boldsymbol{\zeta}^*) + \partial\psi(\mathrm{prox}_\psi \boldsymbol{\zeta}^*) \\
&\Leftrightarrow -\nabla\varphi \circ \mathrm{prox}_\psi(\boldsymbol{\zeta}^*) \in \partial\psi(\mathrm{prox}_\psi \boldsymbol{\zeta}^*) \\
&\Leftrightarrow \boldsymbol{\zeta}^* \in \partial\psi(\mathrm{prox}_\psi \boldsymbol{\zeta}^*) \\
&\Leftrightarrow \boldsymbol{\zeta}^* + \mathrm{prox}_\psi \boldsymbol{\zeta}^* \in (I + \partial\psi)(\mathrm{prox}_\psi \boldsymbol{\zeta}^*) \\
&\Leftrightarrow \mathrm{prox}_\psi \left(\boldsymbol{\zeta}^* + \mathrm{prox}_\psi \boldsymbol{\zeta}^*\right) = \mathrm{prox}_\psi \boldsymbol{\zeta}^*. \qquad (13)
\end{aligned}
$$

Here, $\partial(\varphi + \psi) = \nabla\varphi + \partial\psi$ because $\mathrm{dom}\varphi = \mathbb{R}^n$ due to its differentiability,[4] the third equivalence comes from the assumption, and the final equivalence is due to Fact 1.1. ■

Proposition 2 (A sufficient condition)**.** *A fixed point* $\boldsymbol{\zeta}^* \in \mathbb{R}^n$ *of* T_t *satisfies* $\mathrm{prox}_\psi \boldsymbol{\zeta}^* \in \mathrm{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \varphi(\boldsymbol{w}) + \psi(\boldsymbol{w})$ *if* $\mathrm{prox}_\psi \boldsymbol{\zeta}^* = \mathbf{0}$.

Proof: Clear from the equivalence between (a) and (c) of Theorem 2.2. ■

Example 1.

1) *We consider the case of* $\psi(w) = |w|$, $w \in \mathbb{R}$, *for* $n = 1$. *In this case,* $\mathrm{prox}_\psi \zeta^* = \max\{|\zeta^*| - 1, 0\}\mathrm{sign}(\zeta^*)$. *If* $\mathrm{prox}_\psi \zeta^* \neq 0$, *then* $\zeta^* \notin \partial\psi(\mathrm{prox}_\psi \zeta^*)$ *because* $\zeta^* < -1$ *or* $\zeta^* > 1$ *while* $\partial\psi(\zeta^*) = \{-1\}$ *or* $\partial\psi(\zeta^*) = \{1\}$. *This implies, from Theorem 2, that* $\mathrm{prox}_\psi \zeta^* \notin \mathrm{argmin}_{w\in\mathbb{R}} \varphi(w) + \psi(w)$. *Therefore, together with Proposition 2,* $\mathrm{prox}_\psi \zeta^* = 0$ *is a necessary and sufficient condition to satisfy* $\mathrm{prox}_\psi \zeta^* \in \mathrm{argmin}_{w\in\mathbb{R}} \varphi(w) + \psi(w)$ *in this specific case. It is straightforward to generalize this result to the* ℓ_1 *norm* $\psi(\boldsymbol{w}) = \|\boldsymbol{w}\|_1$ *in a general Euclidean space* $\mathbb{R}^n$*:* $\mathrm{prox}_\psi \boldsymbol{\zeta}^* \in \mathrm{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \varphi(\boldsymbol{w}) + \psi(\boldsymbol{w})$ *if and only if* $\mathrm{prox}_\psi \boldsymbol{\zeta}^* = \mathbf{0}$.
2) *We consider the case of* $\psi(\boldsymbol{w}) = \iota_C(\boldsymbol{w})$, $\boldsymbol{w} \in \mathbb{R}^n$, *for a closed convex set* $C \neq \emptyset$. *In this case,*

$$
\partial\psi(\boldsymbol{w}) = \begin{cases} \left\{\boldsymbol{u} \in \mathbb{R}^n \mid \sup\limits_{\boldsymbol{y}\in C} \langle \boldsymbol{y} - \boldsymbol{w}, \boldsymbol{u}\rangle \leq 0\right\} & \textit{if } \boldsymbol{w} \in C \\ \emptyset & \textit{if } \boldsymbol{w} \notin C \end{cases} \qquad (14)
$$

which is the normal cone to C *at* $\boldsymbol{w}$ [13].

- *When* C *is a closed subspace* M, $\partial\psi(\mathrm{prox}_\psi \boldsymbol{\zeta}^*) = \partial\iota_M(P_M \boldsymbol{\zeta}^*) = M^\perp := \{\boldsymbol{u} \in \mathbb{R}^n \mid \langle \boldsymbol{m}, \boldsymbol{u}\rangle = 0,\ \forall \boldsymbol{m} \in M\}$*; note here that* $P_M \boldsymbol{\zeta}^* \in M$. *Hence, by Theorem 2,* $\mathrm{prox}_\psi \boldsymbol{\zeta}^* \in \mathrm{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \varphi(\boldsymbol{w}) + \psi(\boldsymbol{w})$ *if and only if* $\boldsymbol{\zeta}^* \in M^\perp$ ($\Leftrightarrow \mathrm{prox}_\psi \boldsymbol{\zeta}^* = P_M \boldsymbol{\zeta}^* = \mathbf{0}$)*. This implies that* $\mathrm{prox}_\psi \boldsymbol{\zeta}^* \in \mathrm{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \varphi(\boldsymbol{w}) + \psi(\boldsymbol{w})$ *only in a trivial case.*
- *When* C *is a closed ball* $B := \{\boldsymbol{\zeta} \in \mathbb{R}^n \mid \|\boldsymbol{\zeta}\| \leq \epsilon\}$ *of an arbitrary radius* $\epsilon > 0$,

$$
\partial\psi(\mathrm{prox}_\psi \boldsymbol{\zeta}^*) = \begin{cases} \{\delta\boldsymbol{\zeta}^* \mid \delta \geq 0\} & \textit{if } \boldsymbol{\zeta}^* \notin \mathrm{int}(B), \\ \{\mathbf{0}\}, & \textit{if } \boldsymbol{\zeta}^* \in \mathrm{int}(B), \end{cases} \qquad (15)
$$

where $\mathrm{int}(B)$ *is the interior of the ball* B. *Hence, it holds that* $\mathrm{prox}_\psi \boldsymbol{\zeta}^* \in \mathrm{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \varphi(\boldsymbol{w}) + \psi(\boldsymbol{w})$ *either when* $\boldsymbol{\zeta}^* \in \mathbb{R}^n \setminus \mathrm{int}(B)$ *or when* $\boldsymbol{\zeta}^* = \mathbf{0}$.

We finally present our convergence analysis below.

Theorem 3 (Convergence analysis)**.** *Assume that (i)* $\nabla\varphi$ *is nonexpansive and (ii)* $-\nabla\varphi \circ \mathrm{prox}_\psi$ *has a fixed point. Then, the sequence* $(\boldsymbol{\zeta}_t)_{t\in\mathbb{N}}$ *generated by* (10) *converges to a fixed point* $\boldsymbol{\zeta}^* \in \mathrm{Fix}(-\nabla\varphi \circ \mathrm{prox}_\psi)$, *while* $(\boldsymbol{w}_t)_{t\in\mathbb{N}}$ *converges to* $\mathrm{prox}_\psi \boldsymbol{\zeta}^*$.

Proof: Combining the assumption with Facts 2.1 and 2.2, one can verify that the composition operator $-\nabla\varphi \circ \mathrm{prox}_\psi$ is nonexpansive. Since $\sum_{t=1}^\infty \frac{1}{t}\left(1 - \frac{1}{t}\right) = \sum_{t=1}^\infty \frac{1}{t} - \sum_{t=1}^\infty \left(\frac{1}{t}\right)^2 = \infty$, KM fixed-point theorem [4, 13] (see Theorem 1) can be applied to $(\boldsymbol{\zeta}_t)_{t\in\mathbb{N}}$ to verify the assertion. The convergence of $(\boldsymbol{w}_t)_{t\in\mathbb{N}}$ can be verified by using the nonexpansivity of prox_ψ as $0 \leq \left\|\boldsymbol{w}_t - \mathrm{prox}_\psi \boldsymbol{\zeta}^*\right\| = \left\|\mathrm{prox}_\psi \boldsymbol{\zeta}_t - \mathrm{prox}_\psi \boldsymbol{\zeta}^*\right\| \leq \|\boldsymbol{\zeta}_t - \boldsymbol{\zeta}^*\| \to 0$, $t \to \infty$. ■

IV. Fixed-Point Property of PDA Algorithm under Static Scenario

A. Algorithm Related to PDA and Its Fixed Point Property

We consider the algorithm that generates the sequence $(\boldsymbol{w}_t)_{t\in\mathbb{N}}$ by

$$
\begin{aligned}
\boldsymbol{w}_t &:= \mathrm{prox}_\psi(\boldsymbol{z}_t) \\
\boldsymbol{z}_t &:= T_\varphi(\boldsymbol{z}_{t-1}), \quad \boldsymbol{z}_0 := \mathbf{0}, \qquad (16)
\end{aligned}
$$

where

$$
T_\varphi := I - \eta\nabla\varphi \circ \mathrm{prox}_\psi, \quad \eta > 0. \qquad (17)
$$

[4] In the finite dimensional case, a sufficient condition for having $\partial(\varphi+\psi) = \partial\varphi + \partial\psi$ is $\mathbf{0} \in \mathrm{int}(\mathrm{dom}\psi - \mathrm{dom}\varphi)$, In the present case, $\mathrm{dom}\varphi = \mathbb{R}^n$ and $\mathrm{dom}\psi \neq \emptyset$ so that $\mathrm{int}(\mathrm{dom}\psi - \mathrm{dom}\varphi) = \mathrm{dom}\psi - \mathrm{dom}\varphi = \mathbb{R}^n$. A weaker sufficient condition [13] is $\mathbf{0} \in \mathrm{ri}(\mathrm{dom}\psi - \mathrm{dom}\ \varphi)$, where $\mathrm{ri}(C) := \{\boldsymbol{x} \in \mathbb{R}^n \mid \mathrm{cone}(C - \boldsymbol{x}) = \mathrm{span}(C - \boldsymbol{x})\}$, where given any set $A \subset \mathbb{R}^n$ $\mathrm{cone}A := \{\alpha\boldsymbol{x} \mid \alpha > 0,\ \boldsymbol{x} \in A\}$ and $\mathrm{span}A := \{\alpha\boldsymbol{x} \mid \alpha \in \mathbb{R},\ \boldsymbol{x} \in A\}$.

This algorithm is closely related to PDA, as shown in the following subsection. This algorithm has the following property:

$$\begin{aligned} T_\varphi(\boldsymbol{z}) = \boldsymbol{z} &\Leftrightarrow \boldsymbol{z} - \eta \nabla\varphi(\mathrm{prox}_\psi \boldsymbol{z}) = \boldsymbol{z} \\ &\Leftrightarrow \nabla\varphi(\mathrm{prox}_\psi \boldsymbol{z}) = \boldsymbol{0} \\ &\Leftrightarrow \mathrm{prox}_\psi \boldsymbol{z} \in \operatorname*{argmin}_{\boldsymbol{y}\in\mathbb{R}^n} \varphi(\boldsymbol{y}). \end{aligned} \tag{18}$$

Suppose that the sequence $(\boldsymbol{z}_t)_{t\in\mathbb{N}}$ converges to some point $\boldsymbol{z} \in \mathbb{R}^n$. In this case, $(\boldsymbol{w}_t)_{t\in\mathbb{N}}$ converges to $\mathrm{prox}_\psi \boldsymbol{z}$ due to the continuity of the operator prox_ψ. Since the limit point $\boldsymbol{z}$ of $(\boldsymbol{z}_t)_{t\in\mathbb{N}}$ will be a fixed point of T_φ (i.e., $T_\varphi(\boldsymbol{z}) = \boldsymbol{z}$ will be satisfied), (18) indicates that the limit point $\mathrm{prox}_\psi \boldsymbol{z}$ of $(\boldsymbol{w}_t)_{t\in\mathbb{N}}$ is a minimizer of the function φ, which is independent of the regularizer ψ. This will be shown by simulation in Section V.

B. Reproduction of PDA Algorithm

Define the specific instantaneous-loss function

$$\varphi_t(\boldsymbol{w}) := \frac{1}{2} d^2(\boldsymbol{w}, C_t), \tag{19}$$

where $C_t (\neq \emptyset)$ is the closed convex set accommodating the information acquired at time instant t. A typical design example for online regression is given as

$$C_t := \left\{ \boldsymbol{w} \in \mathbb{R}^n \mid \boldsymbol{w}^\mathsf{T} \boldsymbol{x}_t = y_t \right\}. \tag{20}$$

In this case, the loss function reduces to the following normalized squared-error:

$$\varphi_t(\boldsymbol{w}) := \frac{(y_t - \boldsymbol{w}^\mathsf{T}\boldsymbol{x}_t)^2}{2\left\|\boldsymbol{x}_t\right\|^2}. \tag{21}$$

For online classification,

$$C_t := \left\{ \boldsymbol{w} \in \mathbb{R}^n \mid y_t \boldsymbol{w}^\mathsf{T} \boldsymbol{x}_t \geq 1 \right\} \tag{22}$$

is typically used, where $y_t \in \{-1, 1\}$; $\boldsymbol{x}_t \neq \boldsymbol{0}$ is assumed implicitly here. The gradient of φ_t at $\boldsymbol{w}_{t-1}$ is given by

$$\nabla\varphi_t(\boldsymbol{w}_{t-1}) = \boldsymbol{w}_{t-1} - P_{C_t}(\boldsymbol{w}_{t-1}). \tag{23}$$

Note here that the firm nonexpansivity of the metric-projection operator P_{C_t} implies the firm nonexpansivity of $\nabla\varphi_t = I - P_{C_t}$ (see Facts 1 and 2.3).

We now consider the algorithm that generates $(\boldsymbol{w}_t)_{t\in\mathbb{N}}$ by

$$\begin{aligned} \boldsymbol{w}_t &:= \mathrm{prox}_\psi(\boldsymbol{z}_t) \\ \boldsymbol{z}_t &:= T_{\varphi_t}(\boldsymbol{z}_{t-1}), \quad \boldsymbol{z}_0 := \boldsymbol{0}, \end{aligned} \tag{24}$$

with φ_t defined in (19). It then follows that

$$\begin{aligned} \boldsymbol{z}_t &= \boldsymbol{z}_{t-1} - \eta\nabla\varphi_t(\mathrm{prox}_\psi \boldsymbol{z}_{t-1}) \\ &= \boldsymbol{z}_{t-1} - \eta\nabla\varphi_t(\boldsymbol{w}_{t-1}) \\ &= -\eta \boldsymbol{s}_t, \end{aligned} \tag{25}$$

where by (23) $\boldsymbol{s}_t = \sum_{\tau=1}^t \nabla\varphi_\tau(\boldsymbol{w}_{\tau-1}) = \sum_{\tau=1}^t \boldsymbol{w}_{\tau-1} - P_{C_\tau}(\boldsymbol{w}_{\tau-1})$. By (24) and (25), we obtain the PDA algorithm $\boldsymbol{w}_t = \mathrm{prox}_\psi(-\eta \boldsymbol{s}_t)$. We remark here that the original PDA algorithm explicitly uses a time-varying metric.

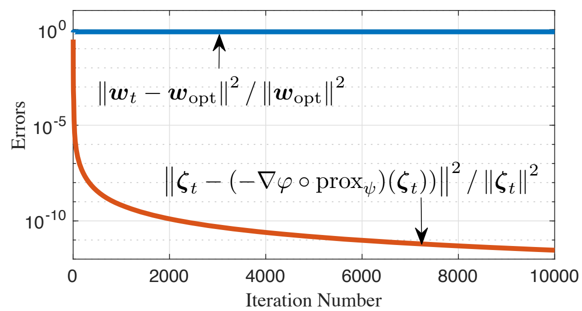

Fig. 1. Simulation results for Theorem 3: $(\boldsymbol{\zeta}_t)_{t\in\mathbb{N}}$ converges to a fixed point $\boldsymbol{\zeta}^*$ of the mapping $-\nabla\varphi \circ \mathrm{prox}_\psi$, but $\boldsymbol{w}^* = \mathrm{prox}_\psi \boldsymbol{\zeta}^*$ is far from the minimizer $\boldsymbol{w}_{\mathrm{opt}}$ of $\varphi + \psi$.

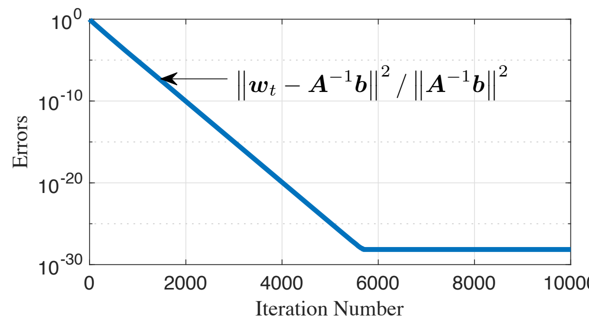

Fig. 2. Simulation results for (18): $(\boldsymbol{w}_t)_{t\in\mathbb{N}}$ converges to the minimizer $\boldsymbol{A}^{-1}\boldsymbol{b} \in \mathrm{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \varphi(\boldsymbol{w})$ of φ.

V. Simulation Studies

We conduct simple simulations to support the theoretical findings of the current work. We consider the quadratic function $\varphi(\boldsymbol{w}) := \frac{1}{2}\left\|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}\right\|^2$ and the regularizer $\psi(\boldsymbol{w}) := 0.1\left\|\boldsymbol{w}\right\|_1$ for $\boldsymbol{w} \in \mathbb{R}^{100}$, where $\boldsymbol{A} := \tilde{\boldsymbol{A}}/\sigma_{\max}(\tilde{\boldsymbol{A}})$. Here, $\sigma_{\max}(\tilde{\boldsymbol{A}})$ is the largest singular value of $\tilde{\boldsymbol{A}}$, and each element of $\tilde{\boldsymbol{A}} \in \mathbb{R}^{100\times 100}$ and $\boldsymbol{b} \in \mathbb{R}^{100}$ are generated randomly from the i.i.d. normal distribution of zero mean and unit variance. The step size for PDA is set to $\eta = 0.1$.

Figure 1 plots the learning curves of two quantities for the RDA algorithm. One is $\left\|\boldsymbol{w}_t - \boldsymbol{w}_{\mathrm{opt}}\right\|^2 / \left\|\boldsymbol{w}_{\mathrm{opt}}\right\|^2$ to see how close the generated solutions are to the optimal point $\boldsymbol{w}_{\mathrm{opt}} \in \mathrm{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \varphi(\boldsymbol{w}) + \psi(\boldsymbol{w})$. Note here that the minimizer exists uniquely due to the strict convexity of φ (due to the full-rankness of $\boldsymbol{A}$) and the coercivity of both φ and ψ. The other quantity is $\left\|\boldsymbol{\zeta}_t - (-\nabla\varphi \circ \mathrm{prox}_\psi)(\boldsymbol{\zeta}_t)\right\|^2 / \left\|\boldsymbol{\zeta}_t\right\|^2$ to illustrate the convergence to a fixed point of $-\nabla\varphi \circ \mathrm{prox}_\psi$. One can see that the second quantity decays, and this is consistent with Theorem 3. Note here that the gradient $\nabla\varphi$ is nonexpansive because $\boldsymbol{A}^\mathsf{T}\boldsymbol{A}$ has a unit spectral norm due to the normalization. We remark that the limit point is not the optimal point $\boldsymbol{w}_{\mathrm{opt}}$, as seen by referring to the first quantity. This is consistent with the arguments in Example 1 (see also Theorem 2).

Figure 2 plots the errors $\left\|\boldsymbol{w}_t - \boldsymbol{A}^{-1}\boldsymbol{b}\right\|^2 / \left\|\boldsymbol{A}^{-1}\boldsymbol{b}\right\|^2$ for the PDA algorithm. One can see that $(\boldsymbol{w}_t)_{t\in\mathbb{N}}$ converges to the minimizer $\boldsymbol{A}^{-1}\boldsymbol{b} \in \operatorname{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \varphi(\boldsymbol{w})$ of φ, which is independent from the regularizer ψ. This is consistent with (18).

VI. Conclusion

We presented the fixed-point theoretic analyses of the RDA and PDA algorithms in the static scenario. If the loss function has a nonexpansive gradient, RDA converges to a fixed point of the mapping $-\nabla\varphi \circ \mathrm{prox}_\psi$ (if exists), and the limit point is a minimizer of the penalized loss function under a restrictive condition. Meanwhile, the fixed point of $I - \eta\nabla\varphi \circ \mathrm{prox}_\psi$ (which is used in PDA implicitly) gives a minimizer of the 'unpenalized' loss function, which is independent from the convex regularizer. The new findings presented in this paper were supported by simulations.

Acknowledgment

This work was supported in part by JSPS Grants-in-Aids (18H01446, 15K13986, 15H02757).

References

[1] Lin Xiao, "Dual averaging methods for regularized stochastic learning and online optimization," *Journal of Machine Learning Research*, vol. 11, pp. 2543–2596, Oct. 2010.

[2] 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.

[3] John Duchi, Elad Hazan, and Yoram Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, Jul. 2011.

[4] 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*, New York, 2011, vol. 49 of *Optimization and Its Applications*, pp. 345–390, Springer.

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

[6] I. Yamada, K. Slavakis, and K. Yamada, "An efficient robust adaptive filtering algorithm based on parallel subgradient projection techniques," *IEEE Trans. Signal Processing*, vol. 50, no. 5, pp. 1091–1101, May 2002.

[7] M. Yukawa and I. Yamada, "Pairwise optimal weight realization — Acceleration technique for set-theoretic adaptive parallel subgradient projection algorithm," *IEEE Trans. Signal Processing*, vol. 54, no. 12, pp. 4557–4571, Dec. 2006.

[8] Masahiro Yukawa, Konstantinos Slavakis, and Isao 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, Feb. 2010.

[9] S. Theodoridis, K. Slavakis, and I. Yamada, "Adaptive learning in a world of projections: a unifying framework for linear and nonlinear classification and regression tasks," *IEEE Signal Process. Mag.*, vol. 28, no. 1, pp. 97–123, Jan. 2011.

[10] A. Ushio and M. Yukawa, "Projection-based dual averaging for stochastic sparse optimization," in *Proc. IEEE ICASSP*, 2017, pp. 2307–2311.

[11] A. Ushio and M. Yukawa, "Projection-based regularized dual averaging for stochastic optimization," *IEEE Trans. Signal Processing*, 2018, submitted for publication.

[12] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *SIAM Journal on Multiscale Modeling and Simulation*, vol. 4, pp. 1168–1200, 2005.

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