AN IMPROVEMENT OF SUBGRADIENT PROJECTION OPERATOR BY COMPOSING MONOTONIC FUNCTIONS

Masao Yamagishi, Isao Yamada

Department of Communications and Integrated Systems
Tokyo Institute of Technology, Tokyo 152-8550, Japan
Email:{myamagi, isao}@comm.ss.titech.ac.jp

ABSTRACT

The subgradient projection operator has been utilized as a computationally efficient tool not only for suppression but also for minimization of convex functions in many applications. In this paper, we propose a systematic scheme to improve significantly the monotone approximation ability, of the subgradient projection, to the level set of a convex function. The proposed scheme is based on a simple observation: the level set of a convex function does not change by composing any zero-crossing monotonically increasing function. A numerical example demonstrates the effectiveness of the proposed scheme in an application to a simple boosting problem.

Index Terms— attracting mapping, monotone approximation operator, subgradient projection, adaptive filtering, machine learning

1. INTRODUCTION

Given a continuous convex function, the subgradient projection operator realizes a monotone approximation to the level set of the function, i.e., the operator can shift any point not in the set strictly closer to the set while it leaves all points in the set unchanged [1, 2, 3]. The subgradient projection operator has been utilized as a computationally efficient tool not only for suppression of convex functions but also for minimization of convex functions in many applications which include for example convex feasibility problems [4, 5, 6], convex optimization problems [2], adaptive filtering problems [1, 3, 7] and machine learning problems [8].

The subgradient projection operator is defined as the metric projection onto a closed half-space defined as the level set of the *tangent plane* of the convex function. The tangent plane can be seen as a simple lower bound touching the function at a point. Obviously only local information on the function, i.e., the value of the function and its subgradient at one specified point, is utilized to define the subgradient projection operator. If certain global information on the function is available additionally, we have seen that much better monotone approximation operators, than the subgradient projection operator, to the level set can be constructed (See for example [9, 10]). Indeed, the operator designs developed in the previous studies [9, 10] essentially rely on special constructions of better lower bounds, than the tangent plane, of the convex function with the aid of such global information. Fortunately, the special construction developed in [9] achieves the best lower bound among all lower bounds which do not exceed any function in a class of certain convex functions. However, if the class is relatively large, the lower bound construction [9] may not achieve a good approximation for some function in the class. On the other hand, another lower bound construction [10] has been developed specially for given a quadratic convex function. This special construction gives a tight lower bound for the quadratic function but is not applicable to other convex functions. So far, there has not yet been established a systematic operator design which is applicable universally to improve sufficiently the monotone approximation ability, of the subgradient projection, to the level set of each convex function even in a wider class. Such a systematic operator design is important especially for certain major convex cost functions including exponential-type functions found for example in boosting techniques [11, 12] for machine learning.

The goal of this paper is to present a systematic operator design which is essentially based on a simple observation: the level set of a convex function does not change by composing any zero-crossing monotonically increasing function. To this goal, we propose a novel operator design scheme which defines a monotone approximation operator as the subgradient projection relative to the composite function designed carefully without losing its convexity. More precisely, the problem of designing a good monotone approximation operator is reduced to that of designing a good zero-crossing monotonically increasing function which does not violate the convexity of the composite function. Fortunately, we found a systematic scheme to design a good zero-crossing monotonically increasing function. By this scheme, we can improve sufficiently the monotone approximation ability, of the subgradient projection, to the level set of a twice differentiable convex function. We also verified the monotone approximation operator found in [10] is reproduced as a special design example of the proposed systematic scheme. Moreover by applying to an exponential-type function, we derived a novel monotone approximation operator which achieves much better approximation performance to the level set than the standard projection relative to the exponential-type function.

Finally, we apply the proposed monotone approximation operator to a simple boosting problem. This application demonstrates the effectiveness of the proposed monotone approximation operators.

2. PRELIMINARIES

Let $\mathbb{R}$ be the set of all real numbers. Let $\mathcal{H}$ be a real Hilbert space equipped with an inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$, $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{H}$, and its induced norm $\|\boldsymbol{x}\| := \langle \boldsymbol{x}, \boldsymbol{x} \rangle^{\frac{1}{2}}$, $\boldsymbol{x} \in \mathcal{H}$. A set $C \subset \mathcal{H}$ is said to be *convex* if $\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y} \in C$ for all $\boldsymbol{x}, \boldsymbol{y} \in C$ and $\alpha \in [0,1]$. For any nonempty closed convex set $C \subset \mathcal{H}$, the metric projection (or simply "projection") $P_C : \mathcal{H} \to C$ maps $\boldsymbol{x} \in \mathcal{H}$ to the unique point $P_C(\boldsymbol{x}) \in C$ such that $\|\boldsymbol{x} - P_C(\boldsymbol{x})\| = \min_{\boldsymbol{y} \in C} \|\boldsymbol{x} - \boldsymbol{y}\|$. A function $f : \mathcal{H} \to \mathbb{R}$ is said to be *convex* if $f(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \le \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{H}$ and $\alpha \in [0,1]$. The *level set* of a function $f : \mathcal{H} \to \mathbb{R}$ is defined by $\mathrm{lev}_{\le 0} f := \{\boldsymbol{x} \in \mathcal{H} : f(\boldsymbol{x}) \le 0\}$. If the function f is continuous convex, then $\mathrm{lev}_{\le 0} f$ is closed convex.

A mapping $T : \mathcal{H} \to \mathcal{H}$ is called a *monotone approximation operator*[9] to a set $C \ne \varnothing$ if T satisfies

$$\begin{cases} \|T(\boldsymbol{x}) - \boldsymbol{y}\| < \|\boldsymbol{x} - \boldsymbol{y}\| & (\forall \boldsymbol{x} \notin C, \ \forall \boldsymbol{y} \in C) \\ T(\boldsymbol{x}) = \boldsymbol{x} & (\forall \boldsymbol{x} \in C). \end{cases}$$

Obviously, $C = \mathrm{Fix}(T) := \{\boldsymbol{x} \in \mathcal{H} \mid T(\boldsymbol{x}) = \boldsymbol{x}\}$ holds. Moreover it is not hard to see that C must be a closed convex set [2, 3] (Note: In *fixed point theory*, the monotone approximation operator is often called an *attracting mapping* [3]).

The fixed point approximation of monotone approximation operators has been applied to many signal processing problems and machine learning problems [1, 3, 7, 8].

The following simple fact can be used to approximate a fixed point of a continuous monotone approximation operator.

Fact 1 ([13]). *Suppose that $\mathcal{H}$ is finite dimensional. Let T be a continuous monotone approximation operator. Then for any initial point $\boldsymbol{x}_0 \in \mathcal{H}$, the sequence $(\boldsymbol{x}_n)_{n\geq 0}$ generated by*

$$\boldsymbol{x}_{n+1} = T\boldsymbol{x}_n \ (n = 0, 1, 2, \ldots)$$

converges to a point in $\mathrm{Fix}(T)$.

3. DESIGN OF BETTER MONOTONE APPROXIMATION OPERATOR TO THE LEVEL SET OF CONVEX FUNCTION

3.1. Key observation for proposed design

In this section, $f\colon \mathcal{H} \to \mathbb{R}$ is assumed to be a twice differentiable and convex, non-constant function with $\mathrm{lev}_{\leq 0} f \neq \varnothing$, hence $\mathcal{R}(f) := \{f(\boldsymbol{x}) \in \mathbb{R} \mid \boldsymbol{x} \in \mathcal{H}\} \ni 0$.

Proposition 1. *Suppose that $h\colon S(\supset \mathcal{R}(f)) \to \mathbb{R}$ is a zero-crossing, i.e., $h(0) = 0$, continuous function which is twice differentiable on $\mathrm{int}\, S$ and $h'(r) > 0, \forall r \in \mathrm{int}\, S$, where $\mathrm{int}\, S$ stands for the interior of S.*

1. *Assume that $h \circ f\colon \mathcal{H} \to \mathbb{R}$ is convex. Then we have*
   - *(i)* $\mathrm{lev}_{\leq 0} f = \mathrm{lev}_{\leq 0}(h \circ f)$;
   - *(ii) $h \circ f$ is continuous on $\mathcal{H}$ and twice differentiable on $\mathcal{D} := \{\boldsymbol{x} \in \mathcal{H} \mid f(\boldsymbol{x}) > \inf_{\boldsymbol{y}\in\mathcal{H}} f(\boldsymbol{y})\}$;*
   - *(iii) The mapping $T_{sp(h\circ f)}\colon \mathcal{H} \to \mathcal{H}$,*

$$T_{sp(h\circ f)}(\boldsymbol{z}) \quad := \quad \begin{cases} \boldsymbol{z} - \frac{h(f(\boldsymbol{z}))}{h'(f(\boldsymbol{z}))}\frac{f'(\boldsymbol{z})}{\|f'(\boldsymbol{z})\|^2} & \text{if } f(\boldsymbol{z}) > 0 \\ \boldsymbol{z} & \text{otherwise} \end{cases}$$

is defined as the projection onto

$$\Lambda^-_{h,f}(\boldsymbol{z}) := \begin{cases} \{\boldsymbol{x} \in \mathcal{H} \mid \langle \boldsymbol{x} - \boldsymbol{z}, (h\circ f)'(\boldsymbol{z})\rangle \\ \qquad\qquad + (h\circ f)(\boldsymbol{z}) \leq 0\} & \text{if } f(\boldsymbol{z}) > 0 \\ \mathcal{H} & \text{otherwise} \end{cases}$$
$$\supset \quad \mathrm{lev}_{\leq 0} f$$

is continuous and 1-strongly attracting (see [2, 3]), i.e.,

$$\left.\begin{array}{l} \|\boldsymbol{z} - T_{sp(h\circ f)}(\boldsymbol{z})\|^2 \leq \|\boldsymbol{z} - \boldsymbol{y}\|^2 - \|T_{sp(h\circ f)}(\boldsymbol{z}) - \boldsymbol{y}\|^2 \\ (\forall \boldsymbol{z} \in \mathcal{H}, \forall \boldsymbol{y} \in \mathrm{Fix}(T_{sp(h\circ f)}) = \mathrm{lev}_{\leq 0} f). \end{array}\right\} \quad (1)$$

2. *$h \circ f$ is convex if and only if*

$$\frac{\langle (f''(\boldsymbol{x}))(\boldsymbol{y}), \boldsymbol{y}\rangle}{\langle f'(\boldsymbol{x}), \boldsymbol{y}\rangle^2} \geq -\frac{h''(f(\boldsymbol{x}))}{h'(f(\boldsymbol{x}))} \qquad (2)$$

for any $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D} \times \mathcal{H}$ such that $\langle f'(\boldsymbol{x}), \boldsymbol{y}\rangle \neq 0$.

Remark 1. *1. By (1), $T_{sp(h\circ f)}$ is a monotone approximation operator to* $\mathrm{lev}_{\leq 0} f$.

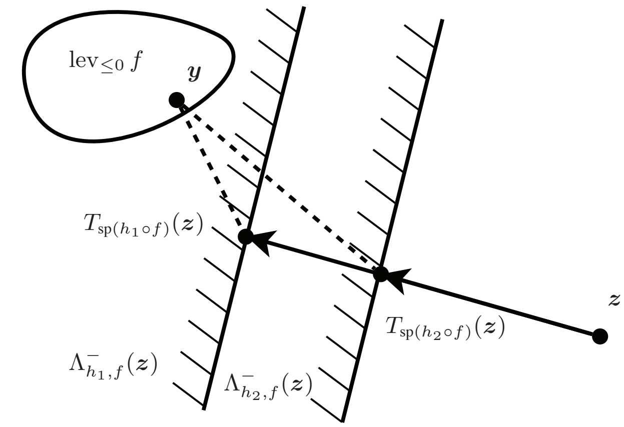


Fig. 1. The operator $T_{\mathrm{sp}(h_1\circ f)}$ is better than $T_{\mathrm{sp}(h_2\circ f)}$ if (4) holds, because, for any $\boldsymbol{z} \notin \mathrm{lev}_{\leq 0} f$, a point $T_{\mathrm{sp}(h_1\circ f)}(\boldsymbol{z})$ is closer to every $\boldsymbol{y} \in \mathrm{lev}_{\leq 0} f$ than a point $T_{\mathrm{sp}(h_2\circ f)}(\boldsymbol{z})$ (see Theorem 1). $T_{\mathrm{sp}(h_i\circ f)}(\boldsymbol{z})$ is the projection onto $\Lambda^-_{h_i,f}(\boldsymbol{z})$ for each $\boldsymbol{z} \in \mathcal{H}$ (i=1,2) (see Proposition 1).

2. *The function*

$$h_0\colon S \to \mathbb{R},\ r \mapsto r,$$

satisfies all requirements on h in Proposition 1-1 and the standard subgradient projection $T_{sp(f)}$ [1, 2, 3] is given by $T_{sp(f)} := T_{sp(h_0\circ f)}$.

In general, there are many functions satisfying all requirements on h in Proposition 1-1. What is a good function h in Proposition 1 in order to approximate $\mathrm{lev}_{\leq 0} f$?

The next theorem suggests a criterion for choosing better $h \in H_f$.

Theorem 1. *Given $S \supset \mathcal{R}(f)$, let*

$$H_f := \left\{ \hat{h}\colon S \to \mathbb{R} \;\middle|\; \begin{array}{c} \hat{h} \text{ satisfies all requirements} \\ \text{on } h \text{ in Proposition 1-1,} \end{array} \right\} \qquad (3)$$

which is nonempty (this is because $h_0 \in H_f$). Suppose that $h_1, h_2 \in H_f$ satisfy

$$-\frac{h_1''(f(\boldsymbol{x}))}{h_1'(f(\boldsymbol{x}))} \geq -\frac{h_2''(f(\boldsymbol{x}))}{h_2'(f(\boldsymbol{x}))},\ \forall \boldsymbol{x} \in \mathcal{D}. \qquad (4)$$

Then we have

$$\|T_{sp(h_1\circ f)}(\boldsymbol{z}) - \boldsymbol{y}\| \leq \|T_{sp(h_2\circ f)}(\boldsymbol{z}) - \boldsymbol{y}\|$$
$$(\forall \boldsymbol{z} \notin \mathrm{lev}_{\leq 0} f, \forall \boldsymbol{y} \in \mathrm{lev}_{\leq 0} f).$$

Theorem 1 tells us that a function $h \in H_f$ yielding larger $-h''(f(\boldsymbol{x}))/h'(f(\boldsymbol{x}))$ realizes better $T_{\mathrm{sp}(h\circ f)}$ for approximating to every point in $\mathrm{lev}_{\leq 0} f$ (see Figure 1).

Thanks to Fact 1, we have the following proposition for application of $T_{\mathrm{sp}(h\circ f)}$ in Proposition 1.

Proposition 2. *Suppose that $\mathcal{H}$ is finite dimensional. Let $f_i\colon \mathcal{H} \to \mathbb{R}$ be a twice differentiable, convex and non-constant function ($i = 1, \ldots, m$), as well as $\cap_{i=1}^m \mathrm{lev}_{\leq 0} f_i \neq \varnothing$. Moreover, for each f_i, let $S_i \supset R(f_i)$ and H_{f_i} be as defined in (3). For arbitrarily fixed $\lambda_i \in (0, 2)$ and any selection $h_i \in H_{f_i}$, define $T_i := I + \lambda_i(T_{sp(h_i\circ f_i)} - I)$ ($i = 1, 2, \ldots, m$), where $I\colon \mathcal{H} \to \mathcal{H}$ stands for the identity* *mapping, i.e.,* $I(\boldsymbol{x}) = \boldsymbol{x}, \forall \boldsymbol{x} \in \mathcal{H}$. *Then for any initial point* $\boldsymbol{x}_0 \in \mathcal{H}$, *the sequence* $(\boldsymbol{x}_n)_{n\geq 0}$ *generated by*

$$\boldsymbol{x}_{n+1} = T\boldsymbol{x}_n \ (n = 0, 1, 2, \ldots), \tag{5}$$

where $T := T_1T_2\ldots T_m$, *converges to a point in* $\mathrm{Fix}(T) = \cap_{i=1}^{m} \mathrm{lev}_{\leq 0} f_i$.

3.2. Examples of proposed monotone approximation operators

In this subsection and next section, we set $\mathcal{H} = \mathbb{R}^n$ equipped with an inner product $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \boldsymbol{x}^T\boldsymbol{y}$, $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{H}$, and its induced norm $\|\boldsymbol{x}\| := \langle \boldsymbol{x}, \boldsymbol{x}\rangle^{\frac{1}{2}}$, $\boldsymbol{x} \in \mathcal{H}$.

In Example 1, we present a good monotone approximation operator for a quadratic function which has been widely used in recent adaptive filtering problems [1]. Part of this result is a reproduction of the monotone approximation operator introduced in [10] as a special case of the use of Theorem 1.

Example 1. *Given* $A \in \mathbb{R}^{m\times n}, \boldsymbol{b} \in \mathbb{R}^m$ *and* $\rho \in \mathbb{R}$, *suppose*

$$f\colon \mathbb{R}^n \to \mathbb{R},\ \boldsymbol{x} \mapsto \|A\boldsymbol{x} - \boldsymbol{b}\|^2 - \rho$$

satisfies $\mathrm{lev}_{\leq 0} f \neq \varnothing$ *and* $\boldsymbol{x}_* := (A^TA)^{-1}A^T\boldsymbol{b}$ *is well defined. Hence* $f'(\boldsymbol{x}_*) = 0$ *holds. Let* $\xi \leq \min_{\boldsymbol{y}\in\mathbb{R}^n} f(\boldsymbol{y})$, $S := \{r \in \mathbb{R} \mid r \geq \xi\}(\supset \mathcal{R}(f))$, *and define*

$$h_{(Q,\xi)}\colon S \to \mathbb{R},\ r \mapsto \sqrt{r-\xi} - \sqrt{-\xi}.$$

Then $h_{(Q,\xi)}$ *satisfies not only all requirements on* h *in Proposition 1-1 but also*

$$\frac{\langle (f''(\boldsymbol{x}))(\boldsymbol{y}), \boldsymbol{y}\rangle}{\langle f'(\boldsymbol{x}), \boldsymbol{y}\rangle^2} \geq \frac{1}{2}\frac{1}{f(\boldsymbol{x}) - \xi} = -\frac{h''_{(Q,\xi)}(f(\boldsymbol{x}))}{h'_{(Q,\xi)}(f(\boldsymbol{x}))} \geq -\frac{h''_0(f(\boldsymbol{x}))}{h'_0(f(\boldsymbol{x}))},$$

$\forall(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}\times\mathcal{H}$ *(see* (2) *and* (4)*). Hence the mapping* $T_{sp(h_{(Q,\xi)}\circ f)}\colon \mathbb{R}^n \to \mathbb{R}^n$,

$$T_{sp(h_{(Q,\xi)}\circ f)}(\boldsymbol{z}) = \begin{cases} \boldsymbol{z} - 2(f(\boldsymbol{z}) - \xi - \sqrt{f(\boldsymbol{z}) - \xi}\sqrt{-\xi})\frac{f'(\boldsymbol{z})}{\|f'(\boldsymbol{z})\|^2} \\ \hfill \text{if } f(\boldsymbol{z}) > 0 \\ \boldsymbol{z} \hfill \text{otherwise,} \end{cases}$$

is a continuous monotone approximation operator to $\mathrm{lev}_{\leq 0} f$ *and satisfies*

$$\|T_{sp(h_{(Q,\xi)}\circ f)}(\boldsymbol{z}) - \boldsymbol{y}\| \leq \|T_{sp(h_0\circ f)}(\boldsymbol{z}) - \boldsymbol{y}\| < \|\boldsymbol{z} - \boldsymbol{y}\|$$

$$(\forall \boldsymbol{z} \notin \mathrm{lev}_{\leq 0} f, \forall \boldsymbol{y} \in \mathrm{lev}_{\leq 0} f).$$

Moreover, if we specially set $\xi = \min_{\boldsymbol{y}\in\mathbb{R}^n} f(\boldsymbol{y})$*(see [10]), then* $h_{(Q,\xi)}$ *is the best design in* H_f, *i.e.,*

$$\|T_{sp(h_{(Q,\xi)}\circ f)}(\boldsymbol{z}) - \boldsymbol{y}\| \leq \|T_{sp(h\circ f)}(\boldsymbol{z}) - \boldsymbol{y}\| < \|\boldsymbol{z} - \boldsymbol{y}\|$$

$$(\forall \boldsymbol{z} \notin \mathrm{lev}_{\leq 0} f, \forall \boldsymbol{y} \in \mathrm{lev}_{\leq 0} f, \forall h \in H_f).$$

In Example 2, we present a good monotone approximation operator to the level set of an exponential-type function which has been used widely in machine learning problems.

Example 2. *Given* $\boldsymbol{a}_i \in \mathbb{R}^n (i = 1, \ldots, m)$ *and* $\rho \in \mathbb{R}$, *suppose*

$$f\colon \mathbb{R}^n \to \mathbb{R}, \boldsymbol{x} \mapsto \sum_{i=1}^{m} \exp(\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle) - \rho \tag{6}$$

satisfies $\mathrm{lev}_{\leq 0} f \neq \varnothing$. *Let* $S := \{r \in \mathbb{R} | r > -\rho\}(\supset \mathcal{R}(f))$ *and define*

$$h_E\colon S \to \mathbb{R},\ r \mapsto \log(r + \rho) - \log(\rho). \tag{7}$$

Then h_E *satisfies not only all requirements on* h *in Proposition 1-1 but also*

$$\begin{aligned} \frac{\langle (f''(\boldsymbol{x}))(\boldsymbol{y}), \boldsymbol{y}\rangle}{\langle f'(\boldsymbol{x}), \boldsymbol{y}\rangle^2} &> \frac{1}{f(\boldsymbol{x}) + \rho} = -\frac{h''_E(f(\boldsymbol{x}))}{h'_E(f(\boldsymbol{x}))} \\ &\geq -\frac{h''_{(Q,-\rho)}(f(\boldsymbol{x}))}{h'_{(Q,-\rho)}(f(\boldsymbol{x}))} \geq -\frac{h''_0(f(\boldsymbol{x}))}{h'_0(f(\boldsymbol{x}))}, \end{aligned}$$

$\forall(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}\times\mathcal{H}$ *(see* (2) *and* (4)*). Hence the mapping* $T_{sp(h_E\circ f)}\colon \mathbb{R}^n \to \mathbb{R}^n$,

$$T_{sp(h_E\circ f)}(\boldsymbol{z}) = \begin{cases} \boldsymbol{z} - (f(\boldsymbol{z}) + \rho)\log(\frac{f(\boldsymbol{z})+\rho}{\rho})\frac{f'(\boldsymbol{z})}{\|f'(\boldsymbol{z})\|^2} & \text{if } f(\boldsymbol{z}) > 0 \\ \boldsymbol{z} & \text{otherwise,} \end{cases}$$

is a continuous monotone approximation operator to $\mathrm{lev}_{\leq 0} f$ *and satisfies*

$$\begin{aligned} \|T_{sp(h_E\circ f)}(\boldsymbol{z}) - \boldsymbol{y}\| &\leq \|T_{sp(h_{(Q,-\rho)}\circ f)}(\boldsymbol{z}) - \boldsymbol{y}\| \\ &\leq \|T_{sp(h_0\circ f)}(\boldsymbol{z}) - \boldsymbol{y}\| < \|\boldsymbol{z} - \boldsymbol{y}\| \\ &\qquad (\forall \boldsymbol{z} \notin \mathrm{lev}_{\leq 0} f, \forall \boldsymbol{y} \in \mathrm{lev}_{\leq 0} f). \end{aligned}$$

4. APPLICATION TO BOOSTING

Boosting is a way to realize a highly accurate classification rule by combining many simple classification rules (base classifiers) into a single. One of the most successful boosting algorithms is Adaboost introduced by Freund and Schapire [11]. This algorithm can also be interpreted as an iterative minimization of the function f defined by (6) [12]. On the other hand, it has been pointed out for example by Grove and Schuurmans [14] that Adaboost may lose generalization performance after many iterations due to excessive suppression of the function f. Therefore an ideal boosting strategy would be fast suppression of f to a certain level with fairly low computational cost. We propose to apply $T_{\mathrm{sp}(h_E\circ f)}$ to the algorithm (5) as an efficient boosting algorithm.

We consider a simple supervised learning problem to identify the unknown classification rule:

$$l\colon [-1,1]^4 \to \{-1,1\},\ \boldsymbol{x} \mapsto \begin{cases} 1 & \text{if } \|\boldsymbol{x}\| < 1 \\ -1 & \text{otherwise} \end{cases}$$

by tuning $(w_{j,k})$ in

$$F_{\boldsymbol{w}}(\boldsymbol{x}) := \sum_{j=1}^{4}\sum_{k=1}^{40} w_{j,k} f_{j,k}(\boldsymbol{x})$$

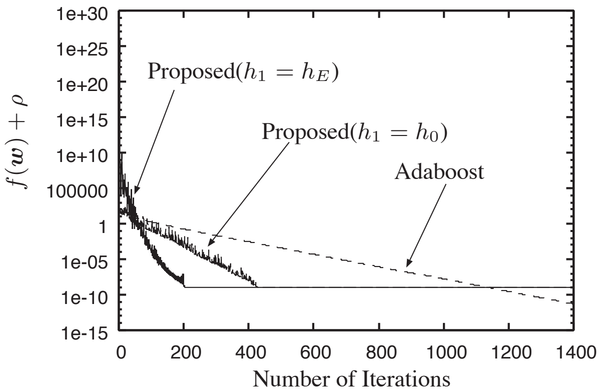
with the knowledge of finite samples $\{(\boldsymbol{x}_i, l(\boldsymbol{x}_i))\}_{i\geq 1}^{50}$, where $\boldsymbol{x}_i \in [-1,1]^4$ $(i = 1, \ldots, 50)$ is drawn from the uniform distribution over $[-1,1]^4$, a class of base classifiers $\{f_{j,k}\colon [-1,1]^4 \to \{-1,1\}$ $(j = 1, \ldots, 4,\ k = 1, \ldots, 40)\}$, are given by

$$f_{j,k}\colon [-1,1]^4 \to \{-1,1\}, \boldsymbol{x} \mapsto \begin{cases} 1 & \text{if } (\boldsymbol{x})_j > \frac{21-k}{20} \\ -1 & \text{otherwise,} \end{cases}$$

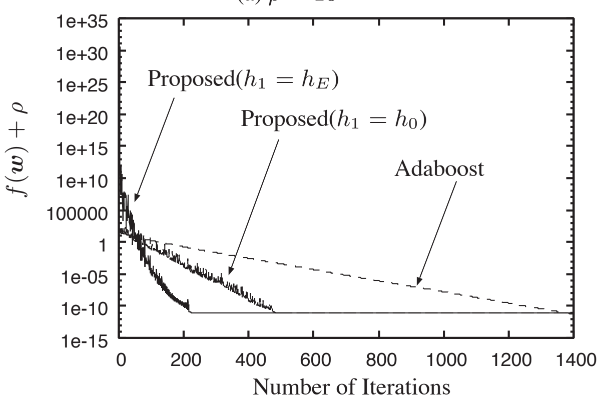
and $(\boldsymbol{x})_j$ is the j-th component of $\boldsymbol{x}$.

For application of the monotone approximation operator $T_{\mathrm{sp}(h_E\circ f)}$ in Example 2 together with the algorithm (5) to this simple boosting problem, we reformulate the above supervised learning problem to the problem of finding a point in the level set of

$$\begin{aligned} f\colon \mathbb{R}^{160} \to \mathbb{R},\ \boldsymbol{w} \quad &\mapsto \quad \sum_{i=1}^{50} \exp\Big(-l(\boldsymbol{x}_i)F_{\boldsymbol{w}}(\boldsymbol{x}_i)\Big) - \rho \\ &=: \quad \sum_{i=1}^{50} \exp(\langle \boldsymbol{a}_i, \boldsymbol{w}\rangle) - \rho. \end{aligned} \tag{8}$$



(a) $\rho = 10^{-9}$



(b) $\rho = 10^{-11}$

Fig. 2. The proposed algorithm (5) for $m = 1$, $f_1 = f$ in (8), $\lambda_1 = 1$ and $h_1 = h_E$ defined by (7) in Example 2 attains the value ρ faster than the others, in each case.

We compared the algorithm (5) with Adaboost which in this scenario can be described equivalently as an iterative minimization of f by the following rule: $\boldsymbol{w}_0 = 0$,

$$\begin{aligned}
\bar{j}_k &= \arg\max_{j=1,\ldots,n} \left| \left(f'(\boldsymbol{w}_k)\right)_j \right|, \ (\boldsymbol{d}_k)_j = \begin{cases} 1 & j = \bar{j}_k \\ 0 & \text{otherwise}, \end{cases} \\
\alpha_k &= \arg\min_{\alpha \in \mathbb{R}} f(\boldsymbol{w}_k + \alpha \boldsymbol{d}_k), \ \boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \alpha_k \boldsymbol{d}_k.
\end{aligned}$$

Figure 2(a) depicts performance comparison between Adaboost and the algorithm (5) where we used $m = 1$, $f_1 = f$ with $\rho = 10^{-9}$ in (8), $\lambda_1 = 1$. Both algorithms start from common point $\boldsymbol{w}_0 = 0$. Figure 2(b) shows performance of the algorithms in the same setting except for $\rho = 10^{-11}$. Since the proposed algorithm (5) with h_E attains the values ρ faster than the other two algorithms, this example clearly shows notable advantage of the proposed monotone approximation operator $T_{\text{sp}(h_E \circ f)}$ over the standard subgradient projection operator $T_{\text{sp}(h_0 \circ f)} = T_{\text{sp}(f)}$ and Adaboost.

5. CONCLUSION

We proposed a systematic scheme to improve the subgradient projection relative to convex functions. A numerical example demonstrated the effectiveness of the proposed scheme in the application to a simple machine learning problem.

ACKNOWLEDGMENTS

The authors would like to thank Prof. Kohichi Sakaniwa of Tokyo Institute of Technology for helpful discussions and comments. This work was partially supported by JSPS Grants-in-Aid (C-19500186).

REFERENCES

[1] 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, 2002.

[2] I. Yamada and N. Ogura, “Hybrid Steepest Descent Method for Variational Inequality Problem over the Fixed Point Set of Certain Quasi-nonexpansive Mappings,” *Numerical Functional Analysis and Optimization*, vol. 25, no. 7, pp. 619–655, 2004.

[3] D. Mandic and I. Yamada, *Tutorial Textbook: Machine Learning and signal processing applications of fixed point theory*, IEEE ICASSP, 2007.

[4] H. H. Bauschke and J. M. Borwein, “On projection algorithm for solving convex feasibility problems,” *SIAM Review*, vol. 38, pp. 367–426, 1996.

[5] P. L. Combettes, “The Foundation of Set Theoretic Estimation,” *Proceeding of the IEEE*, vol. 81, pp. 182–208, 1993.

[6] P. L. Combettes, “Convex set theoretic image recovery by extrapolated iterations of parallel subgradient projections,” *IEEE Trans. Image Processing*, vol. 6, no. 4, pp. 493–506, 1997.

[7] I. Yamada and N. Ogura, “Adaptive Projected Subgradient Method for Asymptotic Minimization of Sequence of Nonnegative Convex Functions,” *Numerical Functional Analysis and Optimization*, vol. 25, no. 7, pp. 593–617, 2004.

[8] K. Slavakis, S. Theodoridis, and I. Yamada, “Online Kernel-Based Classification Using Adaptive Projection Algorithms,” *IEEE Trans. Signal Processing*, vol. 56, no. 7, pp. 2781–2796, 2008.

[9] M. Yamagishi and I. Yamada, “A Deep Monotone Approximation Operator Based on the Best Quadratic Lower Bound of Convex Functions,” *IEICE Trans. Fundamentals*, vol. E91-A, no. 8, pp. 1858–1866, 2008.

[10] N. Ogura and I. Yamada, “A Deep Outer Approximating Half Space of the Level Set of Certain Quadratic Functions,” *Journal of Nonlinear and Convex Analysis*, vol. 6, no. 1, pp. 187–201, 2005.

[11] Y. Freund and R. E. Schapire, “A Decision-Theoretic Generalization of On-Line Learning,” *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[12] J. H. Friedman, T. Hastie, and R. Tibshirani, “Additive logistic regression:A statistical view of boosting,” *Annals of Statistics*, vol. 28, no. 2, pp. 337–407, 2000.

[13] L. Elsner, L. Koltracht, and M. Neumann, “Convergence of sequential and asynchronous nonlinear paracontractions,” *Numerische Mathematik*, vol. 62, pp. 305–319, 1992.

[14] A. J. Grove and D. Schuurmans, “Boosting in the limit: Maximizing the margin of learned ensembles,” *in Proc. of the 15th National / 10th Conference on Artificial Intelligence, American Association for Artificial Intelligence*, pp. 692–699, 1998.