CONVERGENCE RATE ANALYSIS OF PROJECTED STOCHASTIC SUBGRADIENT METHOD USING CONJUGATE GRADIENT-LIKE DIRECTION

YUTA SEKINE AND HIDEAKI IIDUKA

Abstract. This paper considers a stochastic optimization problem of minimizing a strongly convex objective function over a simple, closed convex set and presents a novel projected stochastic subgradient method where only unbiased estimates of subgradients of the objective function are available. The search direction of the proposed method is based on the conjugate gradient directions for unconstrained optimization. We show that the proposed method achieves a convergence rate of $\mathcal{O}(t^{-1})$ under certain assumptions. We also apply the proposed method with the existing method using the steepest descent direction to a support vector machine optimization problem for one data set in the LIBSVM Data and evaluate numerically their methods' performances. The numerical result shows that the proposed method converges to the solution to the problem faster than the existing one and that, compared with the existing method, the classification accuracy is improved by the proposed method.

1. Introduction

In this paper, we consider the following stochastic optimization problem [10, 11]:

Problem 1.1. Suppose that X is a nonempty, bounded, closed convex subset of $\mathbb{R}^n$ with inner product $\langle \cdot, \cdot \rangle$ and its induced norm $\| \cdot \|$, ξ is a random vector whose probability distribution P is supported on set $\Xi \subset \mathbb{R}^n$, and $F \colon X \times \Xi \to \mathbb{R}$. Then

$$\text{minimize } f(w) := \mathbb{E}\left[F\left(w, \xi\right)\right] \text{ subject to } w \in X,$$

where the expectation $\mathbb{E}[F(w, \xi)] := \int_{\Xi} F(w, \xi) dP(\xi)$ is well defined and finite valued for all $w \in X$, and it is assumed that f is Lipschitz continuous and strongly convex.

An interesting example of Problem 1.1 is the support vector machine (SVM) optimization problem [3, 13, 14] defined as follows: given a training set $S := \{(x_m, y_m)\}_{m=1}^{M}$, where $x_m \in \mathbb{R}^{n-1}$ and $y_m \in \{+1, -1\}$,

$$\text{minimize } \frac{\lambda}{2} \|w\|^2 + \mathbb{E}\left[l((w, b); (x, y))\right] \text{ subject to } (w, b) \in X, \tag{1.1}$$

where $X \subset \mathbb{R}^{n-1} \times \mathbb{R}$ is a simple closed convex set (e.g., a closed ball with large enough radius), $\lambda > 0$ is the regularization parameter, the pairs $(x_t, y_t) \in S$ for

2010 *Mathematics Subject Classification.* 90C15, 90C25.

Key words and phrases. conjugate gradient method, projected stochastic subgradient method, stochastic programming.

$t \geq 1$ are independent and identically distributed, and for a given $(x, y) \in \mathbb{R}^{n-1} \times \mathbb{R}$, $l((w, b); (x, y)) := \max\{0, 1 - y(\langle w, x \rangle + b)\}$ $((w, b) \in \mathbb{R}^{n-1} \times \mathbb{R})$.

The classical method for solving Problem 1.1 is the projected stochastic subgradient method [2, (5.4.1)], [10, (1)], [11, (2.1)] defined as follows: given $w_0 \in \mathbb{R}^n$ and $(\gamma_t)_{t \geq 1} \subset [0, +\infty)$,

$$w_t := P_X \left(w_{t-1} - \gamma_t \mathsf{G}(w_{t-1}, \xi_{t-1}) \right) \ (t \geq 1), \tag{1.2}$$

where P_X stands for the metric projection onto X, and it is assumed that (i) there is an independent identically distributed sample $\xi_0, \xi_1, \ldots$ of realizations of random vector ξ and that (ii) there is an oracle which, for a given input point $(w, \xi) \in X \times \Xi$, returns a stochastic subgradient $\mathsf{G}(w, \xi)$ such that $\mathsf{g}(w) := \mathbb{E}[\mathsf{G}(w, \xi)]$ is well defined and is a subgradient of f at w (i.e., $\mathsf{g}(w) \in \partial f(w)$) [10, (b)], [11, (A1), (A2)]. The previously reported results [10, Section 3.2], [11, (2.9))] show that Algorithm (1.2) with $\gamma_t := \theta/(t+1)$ satisfies that, for all $t \geq 1$,

$$\mathbb{E}\left[\|w_t - w^\star\|^2 \right] = \mathcal{O}\left(\frac{1}{t} \right), \tag{1.3}$$

where $\theta > 0$ is a constant depending on the strong convexity constant of f and $w^\star$ is the unique solution to Problem 1.1.

Meanwhile, the *conjugate gradient methods* [12, Chapter 5] are the most popular methods that can accelerate the steepest descent method for a problem of minimizing a smooth function $h \colon \mathbb{R}^n \to \mathbb{R}$ over a whole space $\mathbb{R}^n$. The search direction of the conjugate gradient method is defined for all $t \geq 1$ by

$$d_t := -\nabla h \left(w_{t-1} \right) + \beta_t d_{t-1}, \tag{1.4}$$

where w_{t-1} is the $(t-1)$th approximation to the problem, d_{t-1} is the $(t-1)$th search direction, ∇h stands for the gradient of h, and $\beta_t \geq 0$ $(t \geq 1)$. The well-known formulas of β_t [12, Chapter 5] are, for example, the Fletcher–Reeves, Polak–Ribiére–Polyak, Hestenes–Stiefel, and Dai–Yuan formulas. The formulas do not always satisfy $\lim_{t \to \infty} \beta_t = 0$.

There are some algorithms [5, 6, 7, 8, 9] using (1.4) with $\lim_{t \to \infty} \beta_t = 0$ for deterministic constrained convex optimization. To distinguish between the conventional conjugate gradient directions with the four formulas and the direction with $\lim_{t \to \infty} \beta_t = 0$, we call the latter the *conjugate gradient-like direction*. The previously reported results [5, 6, 7, 8] showed that the gradient algorithms with the conjugate gradient-like directions converge faster than the algorithm [15] with the steepest descent direction.

From the above discussion, we can present the following novel stochastic subgradient method using (1.2) and (1.4), where $\nabla h(w_{t-1})$ is replaced with $-\mathsf{G}(w_{t-1}, \xi_{t-1})$, for solving Problem 1.1: given $w_0 \in \mathbb{R}^n$ and $\mathsf{d}_0 := -\mathsf{G}\left(w_0, \xi_0\right)$,

$$\begin{aligned} \mathsf{d}_t &:= -\mathsf{G}\left(w_{t-1}, \xi_{t-1}\right) + \beta_t \mathsf{d}_{t-1}, \\ w_t &:= P_X \left(w_{t-1} + \gamma_t \mathsf{d}_t \right) \ (t \geq 1). \end{aligned} \tag{1.5}$$

The proposed method with $\beta_t := 0$ $(t \geq 1)$ coincides with the classical projected stochastic subgradient method (1.2). In this paper, we show that the proposed method achieves a convergence rate of $\mathcal{O}(t^{-1})$ under certain assumptions (Theorem 3.1). Finally, we apply the existing and proposed methods to the SVM optimization

problem for the LIBSVM Data [3] and compare numerically the existing method (1.2) [10, 11] with the proposed method (1.5). The numerical result demonstrates that the proposed method performed better than the existing one.

This paper is organized as follows. Section 2 gives the mathematical preliminaries. Section 3 presents the proposed projected stochastic subgradient method for solving the main problem and describes its convergence rate. Section 4 presents numerical evaluation using the LIBSVM Data [3] and compares the behaviors of the existing and proposed methods. Section 5 concludes the paper with a brief summary.

2. Mathematical preliminaries

Let $\mathbb{R}^n$ be an n-dimensional Euclidean space with inner product $\langle\cdot,\cdot\rangle$ and its induced norm $\|\cdot\|$. We denote the history of the process $\xi_0, \xi_1, \ldots$ up to time t by $\xi_{[t]} = (\xi_0, \xi_1, \ldots, \xi_t)$. Unless stated otherwise, all relations between random variables are supported to hold almost surely.

The metric projection [1, Subchapter 4.2, Chapter 28] onto a nonempty, closed convex set $X \subset \mathbb{R}^n$, denoted by P_X, is defined for all $x \in \mathbb{R}^n$ by $P_X(x) \in X$ and $\|x - P_X(x)\| = \inf_{y\in X} \|x - y\|$. P_X is nonexpansive, i.e., $\|P_X(x) - P_X(y)\| \leq \|x - y\|$ for all $x, y \in \mathbb{R}^n$, and $\mathrm{Fix}(P_X) := \{x \in \mathbb{R}^n \colon P_X(x) = x\} = C$ [1, Proposition 4.8, (4.8)].

Let $c > 0$. A function $f \colon \mathbb{R}^n \to \mathbb{R}$ is said to be *c-strongly convex* [1, Definition 10.5] if, for all $x, y \in \mathbb{R}^n$ and for all $\alpha \in (0, 1)$, $f(\alpha x + (1-\alpha)y) + (c\alpha(1-\alpha)/2)\|x - y\|^2 \leq \alpha f(x) + (1-\alpha)f(y)$. The *subdifferential* [1, Definition 16.1] of $f \colon \mathbb{R}^n \to \mathbb{R}$ is defined for all $x \in \mathbb{R}^n$ by

$$\partial f(x) := \{u \in H \colon f(y) \geq f(x) + \langle y - x, u\rangle \ \ (y \in \mathbb{R}^n)\}.$$

We call $u \in \partial f(x)$ the *subgradient* of f at $x \in \mathbb{R}^n$. If f is c-strongly convex, ∂f is strongly monotone; i.e., $\langle x-y, u-v\rangle \geq c\|x-y\|^2$ $(x, y \in \mathbb{R}^n, u \in \partial f(x), v \in \partial f(y))$ [1, Example 22.3(iv)]. Let $L > 0$. $f \colon \mathbb{R}^n \to \mathbb{R}$ is *L-Lipschitz continuous* [1, Definition 1.46] if $|f(x) - f(y)| \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^n$. f is *locally Lipschitz continuous* near a point $x \in \mathbb{R}^n$ if there exists $\rho > 0$ such that $f|_{B(x;\rho)}$ is Lipschitz continuous, where $B(x; \rho)$ stands for closed ball with center x and radius ρ. f is L-locally Lipschitz continuous on a subset $X \subset \mathbb{R}^n$ [1, Definition 1.46] if it is locally Lipschitz continuous near every point in X.

Proposition 2.1. [4, Proposition 3.1(ii)] *Let $f \colon \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz continuous on a nonempty, closed convex set $X \subset \mathbb{R}^n$. Then, f is c-strongly convex if and only if, for all $x, y \in X$ and for all $u \in \partial f(y)$, $f(x) - f(y) \geq \langle u, x - y\rangle + c\|x - y\|^2$.*

3. Convergence rate analysis of proposed method

The following establishes the rate of convergence of the proposed method (1.5) for solving Problem 1.1.

Theorem 3.1. *Assume that the conditions in Problem 1.1, (i), and (ii) hold and (iii) there exists a positive number B such that $\mathbb{E}[\|\mathsf{G}(w, \xi)\|^2] \leq B^2$ for all $(w, \xi) \in X \times \Xi$. Then the sequence $(w_t)_{t\geq 0}$ generated by (1.5) satisfies the following:*

(a) *If* $\gamma_t := 1/(ct)$ *and* $\beta_t \leq 1/t$ *for all* $t \geq 1$*, then, for all* $t \geq 1$*,*

$$\mathbb{E}\left[\|w_t - w^\star\|^2\right] = \mathcal{O}\left(\frac{1+\log t}{t}\right).$$

(b) *If* $\gamma_t := 2/(c(t+1))$ *and* $\beta_t \leq 1/t$ *for all* $t \geq 1$*, then, for all* $t \geq 1$*,*

$$\mathbb{E}\left[\|w_t - w^\star\|^2\right] = \mathcal{O}\left(\frac{1}{t}\right).$$

Proof. We first shaw that $(\mathbb{E}[\|\mathsf{d}_t\|])_{t\geq 0}$ is bounded. Jensen's inequality and (iii) ensure that $\mathbb{E}[\|\mathsf{G}(w,\xi)\|] \leq B$ for all $(w,\xi) \in X \times \Xi$, which implies that $\mathbb{E}[\|\mathsf{G}(w_t,\xi_t)\|] \leq B$ for all $t \geq 0$. The condition $\beta_t \leq 1/t$ $(t \geq 1)$ means $\lim_{t\to\infty} \beta_t = 0$. Accordingly, there exists $t_0 \geq 1$ such that, for all $t \geq t_0$, $\beta_t \leq 1/2$. Putting $R := \max\{B, \mathbb{E}[\|\mathsf{d}_0\|]\}$ means that $\mathbb{E}[\|\mathsf{d}_0\|] \leq 2R$. Assume that $\mathbb{E}[\|\mathsf{d}_t\|] \leq 2R$ for some $t \geq t_0$. Then the triangle inequality leads to the finding that

$$\|\mathsf{d}_{t+1}\| = \|-\mathsf{G}(w_t,\xi_t) + \beta_{t+1}\mathsf{d}_t\| \leq \|\mathsf{G}(w_t,\xi_t)\| + \frac{1}{2}\|\mathsf{d}_t\|.$$

Taking the expectation of both sides of the above inequality implies that

$$\mathbb{E}[\|\mathsf{d}_{t+1}\|] \leq B + \frac{1}{2}\mathbb{E}[\|\mathsf{d}_t\|] \leq 2R.$$

Induction thus shows that $\mathbb{E}[\|\mathsf{d}_t\|] \leq 2R$ for all $t \geq t_0$, i.e., $(\mathbb{E}[\|\mathsf{d}_t\|])_{t\geq 0}$ is bounded.

The nonexpansivity condition of P_X with $P_X(w^\star) = w^\star$ and the equation $\|x+y\|^2 = \|x\|^2 + \|y\|^2 + 2\langle x, y\rangle$ $(x, y \in \mathbb{R}^n)$ imply that, for all $t \geq 1$,

$$\begin{aligned}\|w_t - w^\star\|^2 &= \|P_X(w_{t-1} + \gamma_t\mathsf{d}_t) - P_X(w^\star)\|^2 \\ &\leq \|(w_{t-1} - w^\star) + \gamma_t\mathsf{d}_t\|^2 \\ &= \|w_{t-1} - w^\star\|^2 + \gamma_t^2\|\mathsf{d}_t\|^2 + 2\gamma_t\langle w_{t-1} - w^\star, \mathsf{d}_t\rangle,\end{aligned}$$

which, together with the definition of d_t, implies that

$$\begin{aligned}\|w_t - w^\star\|^2 \leq \|w_{t-1} - w^\star\|^2 + \gamma_t^2\|\mathsf{d}_t\|^2 - 2\gamma_t\langle w_{t-1} - w^\star, \mathsf{G}(w_{t-1},\xi_{t-1})\rangle \\ + 2\gamma_t\beta_t\langle w_{t-1} - w^\star, \mathsf{d}_{t-1}\rangle.\end{aligned} \tag{3.1}$$

Since $w_{t-1} = w_{t-1}(\xi_{[t-2]})$ is independent of ξ_{t-1}, the definition of the expectation ensures that, for all $t \geq 1$,

$$\begin{aligned}\mathbb{E}[\langle w_{t-1} - w^\star, \mathsf{G}(w_{t-1},\xi_{t-1})\rangle] &= \mathbb{E}\left[\mathbb{E}\left[\langle w_{t-1} - w^\star, \mathsf{G}(w_{t-1},\xi_{t-1})\rangle \,|\xi_{[t-2]}\right]\right] \\ &= \mathbb{E}\left[\left\langle w_{t-1} - w^\star, \mathbb{E}\left[\mathsf{G}(w_{t-1},\xi_{t-1})\,|\xi_{[t-2]}\right]\right\rangle\right] \\ &= \mathbb{E}[\langle w_{t-1} - w^\star, \mathsf{g}(w_{t-1})\rangle].\end{aligned}$$

Proposition 2.1 thus leads to the finding that, for all $t \geq 1$,

$$\mathbb{E}[\langle w_{t-1} - w^\star, \mathsf{G}(w_{t-1},\xi_{t-1})\rangle] \geq \mathbb{E}[f(w_{t-1}) - f(w^\star)] + \frac{c}{2}\mathbb{E}\left[\|w_{t-1} - w^\star\|^2\right]. \tag{3.2}$$

Moreover, the Cauchy-Schwarz inequality means that, for all $t \geq 1$,

$$\mathbb{E}[|\langle w_{t-1} - w^\star, \mathsf{d}_{t-1}\rangle|] \leq \mathbb{E}[\|w_{t-1} - w^\star\|\,\|\mathsf{d}_{t-1}\|],$$

which, together with the boundedness conditions of X and $(\mathbb{E}[\|\mathsf{d}_t\|])_{t\geq 1}$, implies that there exists $M_1 \in \mathbb{R}$ such that

$$\mathbb{E}\left[|\langle w_{t-1} - w^\star, \mathsf{d}_{t-1}\rangle|\right] \leq M_1. \tag{3.3}$$

From the definition of d_t, for all $t \geq 1$,

$$\begin{aligned}\|\mathsf{d}_t\|^2 &= \|-\mathsf{G}\left(w_{t-1}, \xi_{t-1}\right) + \beta_t \mathsf{d}_{t-1}\|^2 \\ &= \|\mathsf{G}\left(w_{t-1}, \xi_{t-1}\right)\|^2 + \beta_t^2 \|\mathsf{d}_{t-1}\|^2 - 2\beta_t \left\langle \mathsf{G}\left(w_{t-1}, \xi_{t-1}\right), \mathsf{d}_{t-1}\right\rangle.\end{aligned}$$

Accordingly, the boundedness condition of $(\mathbb{E}[\|\mathsf{d}_t\|])_{t\geq 1}$ and (iii) imply that, there exist $M_2, M_3 \in \mathbb{R}$ such that, for all $t \geq 1$,

$$\begin{aligned}\mathbb{E}\left[\|\mathsf{d}_t\|^2\right] &= \mathbb{E}\left[\|\mathsf{G}\left(w_{t-1}, \xi_{t-1}\right)\|^2\right] + \beta_t^2 \mathbb{E}\left[\|\mathsf{d}_{t-1}\|^2\right] - 2\beta_t \mathbb{E}\left[\left\langle \mathsf{G}\left(w_{t-1}, \xi_{t-1}\right), \mathsf{d}_{t-1}\right\rangle\right] \\ &\leq B^2 + M_2\beta_t^2 + M_3\beta_t.\end{aligned} \tag{3.4}$$

Hence, from (3.1), (3.2), (3.3), and (3.4), for all $t \geq 1$,

$$\begin{aligned}\mathbb{E}\left[\|w_t - w^\star\|^2\right] &\leq \mathbb{E}\left[\|w_{t-1} - w^\star\|^2\right] + \gamma_t^2\left(B^2 + M_2\beta_t^2 + \beta_t M_3\right) + 2M_1\gamma_t\beta_t \\ &\quad - 2\gamma_t\left(\mathbb{E}\left[f\left(w_{t-1}\right) - f\left(w^\star\right)\right] + \frac{c}{2}\mathbb{E}\left[\|w_{t-1} - w^\star\|^2\right]\right) \\ &= \left(1 - c\gamma_t\right)\mathbb{E}\left[\|w_{t-1} - w^\star\|^2\right] + \gamma_t^2\left(B^2 + M_2\beta_t^2 + M_3\beta_t\right) \\ &\quad + 2M_1\gamma_t\beta_t - 2\gamma_t\mathbb{E}\left[f\left(w_{t-1}\right) - f\left(w^\star\right)\right],\end{aligned}$$

which implies that, for all $t \geq 1$,

$$\begin{aligned}\mathbb{E}\left[f\left(w_{t-1}\right) - f\left(w^\star\right)\right] &\leq \frac{1 - c\gamma_t}{2\gamma_t}\mathbb{E}\left[\|w_{t-1} - w^\star\|^2\right] - \frac{1}{2\gamma_t}\mathbb{E}\left[\|w_t - w^\star\|^2\right] \\ &\quad + M_1\beta_t + \frac{\gamma_t}{2}\left(B^2 + M_2\beta_t^2 + M_3\beta_t\right).\end{aligned} \tag{3.5}$$

(a) Let $\gamma_t := 1/(ct)$ and $\beta_t \leq 1/t$ for all $t \geq 1$. Then (3.5) leads to the finding that, for all $t \geq 1$,

$$\begin{aligned}\mathbb{E}\left[f\left(w_{t-1}\right) - f\left(w^\star\right)\right] &\leq \frac{c\left(t-1\right)}{2}\mathbb{E}\left[\|w_{t-1} - w^\star\|^2\right] - \frac{ct}{2}\mathbb{E}\left[\|w_t - w^\star\|^2\right] \\ &\quad + \frac{M_1}{t} + \frac{1}{2ct}\left(B^2 + \frac{M_2}{t^2} + \frac{M_3}{t}\right) \\ &\leq \frac{c\left(t-1\right)}{2}\mathbb{E}\left[\|w_{t-1} - w^\star\|^2\right] - \frac{ct}{2}\mathbb{E}\left[\|w_t - w^\star\|^2\right] + \frac{M_4}{t},\end{aligned}$$

where $M_4 := M_1 + (B^2 + M_2 + M_3)/(2c)$. Summing the above inequality from $t = 1$ to $t = T > 0$ implies that, for all $T > 0$,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[f\left(w_{t-1}\right)\right] - f\left(w^\star\right) &\leq \frac{c}{2T}\sum_{t=1}^{T}\left\{(t-1)\,\mathbb{E}\left[\left\|w_{t-1}-w^\star\right\|^2\right] - t\mathbb{E}\left[\left\|w_t - w^\star\right\|^2\right]\right\} \\ &\quad + \frac{M_4}{T}\sum_{t=1}^{T}\frac{1}{t} \\ &\leq -\frac{c}{2}\mathbb{E}\left[\left\|w_T - w^\star\right\|^2\right] + \frac{M_4(1+\log T)}{T},\end{aligned}$$

where the second inequality comes from $\sum_{t=1}^{T}\{(t-1)\mathbb{E}[\|w_{t-1}-w^\star\|^2] - t\mathbb{E}[\|w_t - w^\star\|^2]\} = -T\mathbb{E}[\|w_T - w^\star\|^2]$ and $\sum_{t=1}^{T}(1/t) \leq 1 + \log T$. The convexity of f thus guarantees that, for all T,

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T}w_{t-1}\right)\right] - f\left(w^\star\right) \leq -\frac{c}{2}\mathbb{E}\left[\left\|w_T - w^\star\right\|^2\right] + \frac{M_4(1+\log T)}{T}.$$

Since $w^\star$ is the solution to Problem 1.1 and $(w_t)_{t\geq 1} \subset X$, we have that, for all T,

$$\mathbb{E}\left[\left\|w_T - w^\star\right\|^2\right] \leq \frac{2M_4(1+\log T)}{cT}.$$

(b) Let $\gamma_t := 2/(c(t+1))$ and $\beta_t \leq 1/t$ for all $t \geq 1$. From (3.5), for all $t \geq 1$,

$$\begin{aligned}\mathbb{E}\left[f\left(w_{t-1}\right) - f\left(w^\star\right)\right] &\leq \frac{c(t-1)}{4}\mathbb{E}\left[\left\|w_{t-1}-w^\star\right\|^2\right] - \frac{c(t+1)}{4}\mathbb{E}\left[\left\|w_t - w^\star\right\|^2\right] \\ &\quad + \frac{M_1}{t} + \frac{1}{c(t+1)}\left(B^2 + \frac{M_2}{t^2} + \frac{M_3}{t}\right).\end{aligned}$$

Accordingly, for all $t \geq 1$,

$$\begin{aligned}t\mathbb{E}\left[f\left(w_{t-1}\right) - f\left(w^\star\right)\right] &\leq \frac{c(t-1)t}{4}\mathbb{E}\left[\left\|w_{t-1}-w^\star\right\|^2\right] - \frac{ct(t+1)}{4}\mathbb{E}\left[\left\|w_t - w^\star\right\|^2\right] \\ &\quad + M_1 + \frac{t}{c(t+1)}\left(B^2 + \frac{M_2}{t^2} + \frac{M_3}{t}\right) \\ &\leq \frac{c(t-1)t}{4}\mathbb{E}\left[\left\|w_{t-1}-w^\star\right\|^2\right] - \frac{ct(t+1)}{4}\mathbb{E}\left[\left\|w_t - w^\star\right\|^2\right] + M_5,\end{aligned}$$

where $M_5 := M_1 + (B^2 + M_2 + M_3)/c$. Summing up the above inequality from $t = 1$ to $t = T$ ensures that, for all T,

$$\sum_{t=1}^{T}t\mathbb{E}\left[f\left(w_{t-1}\right) - f\left(w^\star\right)\right] \leq -\frac{cT(T+1)}{4}\mathbb{E}\left[\left\|w_T - w^\star\right\|^2\right] + M_5 T,$$

which implies that, for all T,

$$\frac{1}{T(T+1)}\sum_{t=1}^{T}t\mathbb{E}\left[f\left(w_{t-1}\right)\right] - \frac{1}{2}f\left(w^\star\right) \leq -\frac{c}{4}\mathbb{E}\left[\left\|w_T - w^\star\right\|^2\right] + \frac{M_5}{T+1},$$

where $\sum_{t=1}^{T} tf(w^\star) = f(w^\star)(T(T+1))/2$. Therefore, the convexity of f leads to the finding that, for all T,

$$\mathbb{E}\left[f\left(\frac{2}{T(T+1)}\sum_{t=0}^{T-1}(t+1)w_t\right)\right] - f(w^\star) \leq -\frac{c}{2}\mathbb{E}\left[\|w_T - w^\star\|^2\right] + \frac{2M_5}{T+1},$$

which implies that, for all T,

$$\mathbb{E}\left[\|w_T - w^\star\|^2\right] \leq \frac{4M_5}{c(T+1)}.$$

This completes the proof. □

4. Numerical experiment

This section applies the existing method (1.2) and the proposed method (1.5) to the support vector machine optimization problem (1.1) with $\lambda = 1/n$ [10, Section 4] for the data set "a1a" (the number of the learning data is 1605, the number of the unknown data is 30956, and the dimension of the problem is 123) from the LIBSVM Data [3]. The experiment used a 13-inch MacBook Air with Intel(R) Core(TM) i7-5650U CPU processor, 8GB 1600MHz DDR3 RAM memory, and Mac OS X El Capitan (Version 10.11.6) operating system. The algorithms used in the experiment were the following. They were written in MATLAB 2016a (9.0.0.341360).

- Projected Stochastic Subgradient Method (1.2) with $\gamma_t = 2n/(t+1)$ (PSSM)
- Proposed Method (1.5) with $\gamma_t = n/t$ and $\beta_t = 1/t$ (PM1, Theorem 3.1(a))
- Proposed Method (1.5) with $\gamma_t = 2n/(t+1)$ and $\beta_t = 1/t$ (PM2, Theorem 3.1(b))

We set an initial point $w_0 = 0$ in the algorithms and the stopping condition of the algorithms was $t = 10^3$.

The value of $f(w_5)$ generated by each of PSSM, PM1, and PM2 was 2.1074e+04, 1.4424e+04, and 1.5198e+03, respectively. Moreover, we checked that PSSM, PM1, and PM2 converge to the same point at which the value of f is approximately 10^3, i.e., PM2 converges faster than PSSM and PM1.

Table 1 indicates the classification accuracies for PSSM, PM1, and PM2. It shows that, compared with the existing method, the accuracies were improved by the proposed methods.

Table 1. The classification accuracies (%) for the existing method (PSSM) and the proposed methods (PM1 and PM2)

	PSSM	PM1	PM2
a1a [3]	82.15	83.74	83.57

5. Conclusion

This paper presented a novel projected subgradient method for stochastic constrained smooth convex optimization. The proposed method uses the search direction based on the conventional conjugate gradient directions. We showed that

the proposed method achieves a convergence rate of $\mathcal{O}(t^{-1})$ under certain assumptions. Numerical evaluation using a concrete support vector machine optimization problem showed the efficiency of the proposed method.

Acknowledgments

We are sincerely grateful to the editor, Wataru Takahashi, and the anonymous reviewers for helping us improve the original manuscript. We also thank Kazuhiro Hishinuma for his input on the numerical evaluation. This work was supported by the Japan Society for the Promotion of Science through a Grant-in-Aid for Scientific Research (C) (15K04763).

References

[1] H.H. Bauschke and P.L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, New York (2011).

[2] V.S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*, Cambridge University Press, New York (2008).

[3] C.-C. Chang and C.-J. Lin, *LIBSVM–A library for support vector machines*, ACM Trans. Intell. Syst. Tech. **2** (2011), 1–27. URL `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html`

[4] L. Fan, S. Liu, and S. Gao, *Generalized monotonicity and convexity of non-differentiable functions* J. Math. Anal. Appl. **279** (2003), 276–289.

[5] H. Iiduka, *Three-term conjugate gradient method for the convex optimization problem over the fixed point set of a nonexpansive mapping*, Appl. Math. Comput. **217** (2011), 6315–6327.

[6] H. Iiduka, *Fixed point optimization algorithms for distributed optimization in networked systems*, SIAM J. Optim. **23** (2013) 1–26.

[7] H. Iiduka, *Acceleration method for convex optimization over the fixed point set of a nonexpansive mapping*, Math. Program. **149** (2015) 131–165.

[8] H. Iiduka and I. Yamada, *A use of conjugate gradient direction for the convex optimization problem over the fixed point set of a nonexpansive mapping*, SIAM J. Optim. **19** (2009), 1881–1893.

[9] K. Sakurai and H. Iiduka, *Acceleration of the Halpern algorithm to search for a fixed point of a nonexpansive mapping*, Fixed Point Theory Appl. **2014** (2014) 202.

[10] S. Lacoste-Julien, M. Schmidt, and F. Bach, *A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method*, arXiv (2012) URL `https://arxiv.org/pdf/1212.2002.pdf`

[11] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optim. **19** (2009), 1574–1609.

[12] J. Nocedal and S.J. Wright, *Numerical Optimization. Springer Series in Operations Research and Financial Engineering*, Springer, New York (1999).

[13] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter. *Pegasos: primal estimated sub-gradient solver for SVM*, Math. Program. **127** (2011), 3–30.

[14] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. *Large margin methods for structured and interdependent output variables*, J. Mach. Learn. Res. **6** (2006), 1453–1484.

[15] I. Yamada. *The hybrid steepest descent method for the variational inequality problem over the intersection of fixed point sets of nonexpansive mappings*, In: D. Butnariu, Y. Censor, S. Reich (eds.) Inherently Parallel Algorithms for Feasibility and Optimization and Their Applications, pp. 473–504. Elsevier (2001).

(Y. Sekine) Department of Computer Science, Meiji University, 1-1-1 Higashimita, Tama-ku, Kawasaki-shi, Kanagawa 214-8571, Japan
E-mail address: sekine@cs.meiji.ac.jp

(H. Iiduka) Department of Computer Science, Meiji University, 1-1-1 Higashimita, Tama-ku, Kawasaki-shi, Kanagawa 214-8571, Japan
E-mail address: iiduka@cs.meiji.ac.jp