Citation
Yu, T. and Liu, X.W. and Dai, Y.H. and Sun, J. 2022. Variable metric proximal stochastic variance reduced gradient methods for nonconvex nonsmooth optimization. Journal of Industrial and Management Optimization. 18 (4): pp. 2611-2631.
http://doi.org/10.3934/jimo.2021084

JOURNAL OF INDUSTRIAL AND MANAGEMENT OPTIMIZATION
doi:10.3934/jimo.2021084

VARIABLE METRIC PROXIMAL STOCHASTIC VARIANCE REDUCED GRADIENT METHODS FOR NONCONVEX NONSMOOTH OPTIMIZATION

Tengteng Yu

School of Artificial Intelligence
Hebei University of Technology, Tianjin 300401, China

Xin-Wei Liu*

Institute of Mathematics
Hebei University of Technology, Tianjin 300401, China

Yu-Hong Dai

LSEC, Academy of Mathematics and Systems Science
Chinese Academy of Sciences, Beijing 100190, China
School of Mathematical Sciences
University of Chinese Academy of Sciences, Beijing 100049, China

Jie Sun

Institute of Mathematics
Hebei University of Technology, Tianjin 300401, China
School of Business
National University of Singapore, Singapore 119245, Singapore

(Communicated by Kok Lay Teo)

Abstract. We study the problem of minimizing the sum of two functions. The first function is the average of a large number of nonconvex component functions and the second function is a convex (possibly nonsmooth) function that admits a simple proximal mapping. With a diagonal Barzilai-Borwein stepsize for updating the metric, we propose a variable metric proximal stochastic variance reduced gradient method in the mini-batch setting, named VM-SVRG. It is proved that VM-SVRG converges sublinearly to a stationary point in expectation. We further suggest a variant of VM-SVRG to achieve linear convergence rate in expectation for nonconvex problems satisfying the proximal Polyak-Łojasiewicz inequality. The complexity of VM-SVRG is lower than that of the proximal gradient method and proximal stochastic gradient method, and is the same as the proximal stochastic variance reduced gradient method. Numerical experiments are conducted on standard data sets. Comparisons with other advanced proximal stochastic gradient methods show the efficiency of the proposed method.

2020 *Mathematics Subject Classification.* Primary: 90C06; Secondary: 90C30.
Key words and phrases. Nonconvex nonsmooth optimization, proximal stochastic gradient method, Barzilai-Borwein method, variable metric, proximal Polyak-Łojasiewicz inequality.

* Corresponding author: Xin-Wei Liu.

1. **Introduction.** We are interested in the composite minimization problem

$$\min_{w\in\mathbb{R}^d} P(w) = F(w) + R(w), \quad \text{where} \quad F(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w), \tag{1}$$

and each component function $f_i(w) : \mathbb{R}^d \to \mathbb{R}$, $i = 1, 2, \ldots, n$, is smooth and nonconvex, while $R(w) : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a relatively simple convex function but can be nondifferential (referred to as a regularization term). Problems of the form (1) often arise in machine learning [3, 8, 29, 32] and statistics [10], known as regularized empirical risk minimization (ERM).

The proximal gradient descent (Prox-GD) method [19, 22] is popular for solving composite problems. As a generalization of Prox-GD, the variable metric proximal gradient (VM-PG) method [2, 21, 23] can achieve better performance with a proper metric. However, Prox-GD or VM-PG requires to compute the exact full gradient in each iteration, which is computationally prohibitive in the case where n is extremely large. One remedy is the proximal stochastic gradient descent (Prox-SGD) method, which is a variant of stochastic gradient descent (SGD) method that could be dated back to the seminal work of Robbins and Monro [27]. Specifically, in the k-th iteration, Prox-SGD chooses $i_k \in \{1, 2, \ldots, n\}$ uniformly at random and updates the iterate by

$$w_{k+1} = \arg\min_{w\in\mathbb{R}^d} \Big\{ \nabla f_{i_k}(w_k)^T w + \frac{1}{2\eta_k}\|w - w_k\|_2^2 + R(w) \Big\}, \tag{2}$$

where $\nabla f_{i_k}(w_k)$ denotes the gradient of the i_k-th component function f_{i_k} at w_k and $\eta_k > 0$ is the stepsize (a.k.a. learning rate). Given the scaled proximal operator of R relative to the metric A [23]:

$$\mathrm{prox}_R^A(w) = \arg\min_{y\in\mathbb{R}^d} \Big\{ \frac{1}{2}\|y - w\|_A^2 + R(y) \Big\},$$

where $A \in \mathbb{R}^{d\times d}_{++}$ is positive definite and $\|z\|_A = \sqrt{z^T A z}$, the update rule of Prox-SGD in (2) can be equivalently written as

$$w_{k+1} = \mathrm{prox}_R^{\eta_k^{-1}\mathbb{I}_d}\big(w_k - \eta_k \nabla f_{i_k}(w_k)\big), \tag{3}$$

in which $\mathbb{I}_d \in \mathbb{R}^{d\times d}$ is the identity matrix. When $R(w) = 0$, relation (3) becomes the update rule of standard SGD method.

From (2) we see that, at each iteration, Prox-SGD only computes the gradient of a single component function, and thus its computational cost is roughly $1/n$ of that of Prox-GD. Since the random sampling yields a large variance of the stochastic gradient, Prox-SGD only converges sublinearly under strong convexity. Motivated by several prevalent variance-reduced stochastic gradient methods such as stochastic average gradient (SAG) [28], stochastic variance reduced gradient (SVRG) [12], incremental gradient (SAGA) [6], semi-stochastic gradient descent (S2GD) [15], and stochastic recursive gradient (SARAH) [20], many researchers have devoted attention to how to use variance reduction techniques to improve the convergence rate of Prox-SGD. For example, proximal stochastic variance reduction gradient (Prox-SVRG) [34], a mini-batch proximal variant of S2GD (mS2GD) [14] and variance reduced stochastic gradient descent (VR-SGD) [29] achieve linear convergence for the nonsmooth strongly convex or non-strongly convex case. Recently, as the popularization of deep learning, the nonconvex nonsmooth problem (1) has triggered off intensive research work. To mention a few of them, Reddi et al. [26] developed

nonconvex variants of Prox-SVRG and Prox-SAGA, and established their linear convergence under the proximal Polyak-Łojasiewicz (PL) inequality [13]. Li and Li [17] suggested a variant of Prox-SVRG, called Prox-SVRG+, which converges sublinearly to a stationary point and achieves linear rate without restart when the objective function satisfies the PL inequality. A proximal variant of SARAH that converges sublinearly in the nonconvex case has been proposed by Pham et al. [24].

It is well known that the choice of stepsizes has an important influence on SGD both theoretically and numerically [3]. The classical SGD and its proximal variants often employ a diminishing stepsize, or a fixed stepsize tuned by hand. However, these two types of stepsize rules may be time-consuming in practice. In recent years, using the Barzilai-Borwein (BB) method [1] to compute the stepsize has attracted more and more attention in developing efficient SGD methods [18, 30, 35, 36]. For example, Tan et al. [30] suggested to employ the BB method to automatically compute stepsizes for SGD and SVRG, and developed the SGD-BB and SVRG-BB methods. Yu et al. [35] combined the trust-region scheme and BB stepsizes with SARAH for solving nonsmooth convex composite problems. A remarkable advantage of the stepsize given by the BB method is that it estimates a scalar approximation of the Hessian and is not sensitive to the choice of initial stepsizes, see [5, 7, 11] and references therein for more details about BB-like methods. However, the research on incorporating BB stepsizes with proximal stochastic gradient methods in the nonconvex nonsmooth case is far less than in the convex case.

Motivated by the success of the marriage of BB stepsizes and SGD in the convex case, we propose a variable metric proximal stochastic variance reduced gradient method in the mini-batch setting, named VM-SVRG, for solving the nonconvex nonsmooth problem (1). The method employs a diagonal BB stepsize, which is the closed-form solution of a constrained optimization problem and can easily be calculated. Moreover, in each iteration our VM-SVRG method has the same computational cost on gradients as Prox-SVRG. It is proved that VM-SVRG converges sublinearly to a stationary point in expectation. By employing the proximal Polyak-Łojasiewicz (PL) inequality [13, 26], a variant of the VM-SVRG method achieves linear convergence rate in expectation. In addition, the complexity of VM-SVRG is lower than that of Prox-GD and Prox-SGD, and is the same as Prox-SVRG. Numerical experiments on standard data sets including ijcnn1, rcv1, real-sim and covtype show that our proposed VM-SVRG performs better than some advanced mini-batch proximal stochastic gradient methods and their variants including Prox-SVRG, mS2GD, mS2GD-BB, mSARAH (a mini-batch proximal variant of SARAH in [20]), and mSARAH-BB (a mini-batch proximal variant of SARAH-BB in the literature [18]).

The rest of this paper is organized as follows. In Section 2 we propose our VM-SVRG method. In Section 3 we analyze the convergence and complexity of VM-SVRG under different conditions. Numerical experiments are reported in Section 4. Finally, we draw some conclusions in Section 5.

2. The VM-SVRG method.

Notice that Prox-SVRG updates the stochastic gradient as follows

$$v_t^k = \frac{\nabla f_{i_t}(w_t^k) - \nabla f_{i_t}(\tilde{w}_k)}{q_{i_t} n} + \nabla F(\tilde{w}_k),$$

where $i_t \in \{1, 2, \ldots, n\}$ is chosen randomly according to Ω. Such a stochastic gradient provides an unbiased estimate of the full gradient $\nabla F(w_t^k)$. A great advantage

of v_t^k is that its variance is much smaller than that of the stochastic gradient used by Prox-SGD. Moreover, the variance of v_t^k will gradually converge to zero. Consequently, Prox-SVRG with a constant stepsize achieves linear convergence rate as oppose to a sublinear rate of Prox-SGD.

Our VM-SVRG method, presented in Algorithm 1, calculates the stochastic gradient in a mini-batch form of v_t^k used by Prox-SVRG. We mention that v_t^k in VM-SVRG is also an unbiased estimate of the full gradient $\nabla F(w_t^k)$. In fact, conditioned on w_t^k, we take expectation with respect to I_t and obtain

$$\begin{aligned}\mathbb{E}[v_t^k] &= \sum_{i=1}^{n} \frac{\nabla f_i(w_t^k) - \nabla f_i(\tilde{w}_k)}{q_i n} \cdot q_i + \nabla F(\tilde{w}_k) \\ &= \nabla F(w_t^k) - \nabla F(\tilde{w}_k) + \nabla F(\tilde{w}_k) \\ &= \nabla F(w_t^k),\end{aligned}$$

where the second equality follows from the fact $\nabla F(w_t^k) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w_t^k)$.

Algorithm 1 VM-SVRG(w^0, m, b, U_0)

Input: Maximal number of inner iterations m, initial point $\tilde{w}_0 = w^0 \in \mathbb{R}^d$, initial metric U_0, mini-batch size $b \in \{1, 2, \ldots, n\}$, probability $\Omega = \{q_1, q_2, \ldots, q_n\}$;

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Calculate $\tilde{v}_k = \nabla F(\tilde{w}_k)$.
3: Set $w_0^k = \tilde{w}_k$.
4: Choose $t_k \in \{1, 2, \ldots, m\}$ uniformly at random.
5: **for** $t = 0, 1, \ldots, t_k - 1$ **do**
6: Choose mini-batch $I_t \subseteq \{1, 2, \ldots, n\}$ of size b, where each $i \in I_t$ is chosen from $\{1, 2, \ldots, n\}$ randomly according to Ω. Compute

$$v_t^k = \frac{1}{b}\sum_{i \in I_t} \left[\frac{1}{q_i n}(\nabla f_i(w_t^k) - \nabla f_i(\tilde{w}_k))\right] + \tilde{v}_k. \tag{4}$$

7: Compute $w_{t+1}^k = \mathrm{prox}_R^{U_k^{-1}}(w_t^k - U_k v_t^k)$.
8: **end for**
9: Set $\tilde{w}_{k+1} = w_{t_k}^k$.
10: Compute U_k from (6).
11: **end for**

Output: Iterate w_a chosen uniformly at random from $\{\{w_t^k\}_{t=0}^{t_k-1}\}_{k=0}^{K-1}$.

Clearly, when $U_k = \alpha_k \mathbb{I}_d$ with α_k being a scalar stepsize, our VM-SVRG method reduces to mS2GD for $q_i = 1/n$, $i = 1, \ldots, n$, and to Prox-SVRG if we set $b = 1$ and $t_k = m$. Furthermore, if U_k is an approximation of the inverse Hessian, VM-SVRG transforms to a stochastic proximal quasi-Newton method, see [31, 33]. As suggested in [25, 31], w_a is chosen uniformly at random from $\{\{w_t^k\}_{t=0}^{t_k-1}\}_{k=0}^{K-1}$, which can use all the information from both the outer and inner loops.

Notice that the quasi-Newton method captures the second-order information by requiring U_k to satisfy the secant equation $s_k = U_k y_k$ or $U_k s_k = y_k$, where $s_k = \tilde{w}_k - \tilde{w}_{k-1}$, $y_k = \nabla F(\tilde{w}_k) - \nabla F(\tilde{w}_{k-1})$. The first secant equation approximates U_k to the inverse Hessian while the second one approximates it to the Hessian. However, computing a full dense approximation matrix U_k may be extremely expensive in

large-scale setting. Motivated by the diagonal approximation strategy in [23, 37], we suggest to compute U_k as follows

$$\begin{aligned} \min_{u\in\mathbb{R}^d} \quad & \|s_k - Uy_k\|_2^2 + \omega\|U - U_{k-1}\|_F^2 \\ \text{s.t.} \quad & \underline{\alpha}_k\mathbb{I}_d \preceq U \preceq \overline{\alpha}_k\mathbb{I}_d, \\ & U = \mathrm{Diag}(u), \end{aligned} \tag{5}$$

where $\omega > 0$, $\|\cdot\|_F$ is the Frobenius norm and $0 < \underline{\alpha}_k \le \overline{\alpha}_k$ are two stepsizes given by users. In [37], for the convex case of (1), problem (5) is employed to update the metric for a stochastic recursive gradient method and provide very promising results. Apparently, problem (5) provides a solution U_k satisfying the secant equation $s_k = U_k y_k$ in the sense of least squares and ω controls the closeness to the previous metric U_{k-1}. This is different from the one in [23], which constructs U_k by using the secant equation $U_k s_k = y_k$, i.e., replacing the objective in (5) with $\|s_k - Uy_k\|_2^2 + \omega\|U - U_{k-1}\|_F^2$.

An important advantage of problem (5) is that it has a closed-form solution $U_k = \mathrm{Diag}(u_k) \in \mathbb{R}^{d\times d}$ with $u_k = [u_k^{(1)}, u_k^{(2)}, \ldots, u_k^{(d)}]$, where

$$u_k^{(i)} = \begin{cases} \underline{\alpha}_k, & \text{if } \frac{s_k^{(i)}y_k^{(i)} + \omega u_{k-1}^{(i)}}{\left(y_k^{(i)}\right)^2 + \omega} < \underline{\alpha}_k; \\ \overline{\alpha}_k, & \text{if } \frac{s_k^{(i)}y_k^{(i)} + \omega u_{k-1}^{(i)}}{\left(y_k^{(i)}\right)^2 + \omega} > \overline{\alpha}_k; \\ \frac{s_k^{(i)}y_k^{(i)} + \omega u_{k-1}^{(i)}}{\left(y_k^{(i)}\right)^2 + \omega}, & \text{otherwise.} \end{cases} \tag{6}$$

Here $s_k^{(i)}$ and $y_k^{(i)}$ are the i-th elements of s_k and y_k, respectively.

For $\overline{\alpha}_k$ and $\underline{\alpha}_k$, we would like to employ $\alpha_k^D = \frac{\|s_k\|_2}{\|y_k\|_2}$ in [4] and $\alpha_k^{BB} = \frac{s_k^T s_k}{s_k^T y_k}$ in [1], which have been applied in SGD methods, see [30, 35, 36] for example. To avoid negative values of the stepsize and consider unbiased gradient estimators added to w_0^k in the inner loop, we use the following two variants

$$\underline{\alpha}_k = \frac{2b}{m} \cdot \frac{\|s_k\|_2}{\|y_k\|_2}$$

and

$$\overline{\alpha}_k = \frac{2b}{m} \cdot \frac{s_k^T s_k}{|s_k^T y_k|}.$$

Clearly, both $\overline{\alpha}_k$ and $\underline{\alpha}_k$ are nonnegative and $\underline{\alpha}_k \le \overline{\alpha}_k$ always holds due to the Cauchy-Schwarz inequality. To chop extreme values of $\overline{\alpha}_k$ and $\underline{\alpha}_k$, we project them into some interval $[\underline{\alpha}, \overline{\alpha}]$ so that $u_k^{(i)}$ $(i = 1, \ldots, n)$ will be bounded for all k.

3. Convergence analysis.

In our subsequent analysis, we make the following two common assumptions.

Assumption 1. *The function $R(w) : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is proper closed and convex, but can be nondifferentiable.*

Assumption 2. *Each component function $f_i(w) : \mathbb{R}^d \to \mathbb{R}$, for $i = 1, 2, \ldots, n$, is L_i-smooth. That is, there exists $L_i > 0$ such that*

$$\|\nabla f_i(w) - \nabla f_i(v)\|_2 \le L_i\|w - v\|_2, \quad \forall w, v \in \mathbb{R}^d.$$

Denoting $L = \frac{1}{n}\sum_{i=1}^n L_i$, by Assumption 2, we conclude that $F(w) = \frac{1}{n}\sum_{i=1}^n f_i(w)$ is L-smooth and $L_\Omega \geq L$, where

$$L_\Omega = \max_{i=1,2,\ldots,n} \frac{L_i}{nq_i}.$$

We need the notations $u_k^{\max}$ and $u_k^{\min}$ as

$$u_k^{\max} = \max_{j=1,2\ldots,d}\left\{u_k^{(j)}\right\} \quad \text{and} \quad u_k^{\min} = \min_{j=1,2\ldots,d}\left\{u_k^{(j)}\right\}. \tag{7}$$

3.1. **Sublinear convergence.** In this subsection, we establish sublinear convergence of VM-SVRG. We first present some intermediate results.

Lemma 3.1. *(Lemma 8 [37]) Consider $P(w)$ defined in (1). Suppose Assumptions 1 and 2 hold. Let $w' = prox_R^{A^{-1}}(w - A\zeta)$, where $A \in \mathbb{S}_{++}^{d\times d}$ is a symmetric positive definite matrix and $\zeta \in \mathbb{R}^d$. Then,*

$$P(w') \leq P(z) + (w'-z)^T(\nabla F(w) - \zeta) + \frac{1}{2}\|w'-w\|^2_{(L_\Omega \mathbb{I}_d - A^{-1})} + \frac{1}{2}\|z-w\|^2_{(L_\Omega \mathbb{I}_d + A^{-1})} - \frac{1}{2}\|w'-z\|^2_{A^{-1}},\ \forall z \in \mathbb{R}^d.$$

Now we derive an upper bound on variance of v_t^k in the following lemma.

Lemma 3.2. *Let Assumptions 1 and 2 be satisfied, and choose $b \in \{1,2,\ldots,n\}$. Then,*

$$\mathbb{E}[\|v_t^k - \nabla F(w_t^k)\|_2^2] \leq \frac{L_\Omega^2}{b}\|w_t^k - \tilde{w}_k\|_2^2.$$

Proof. Let $I_t = \{i_1, \ldots, i_b\}$ and define

$$\varphi_t^k = \frac{1}{b}\sum_{j=1}^b \varphi_{t,i_j}^k, \text{ where } \varphi_{t,i}^k = \frac{\nabla f_i(w_t^k) - \nabla f_i(\tilde{w}_k)}{q_i n}.$$

Following from the fact $F(w) = \frac{1}{n}\sum_{i=1}^n f_i(w)$, we have

$$\mathbb{E}[\varphi_t^k] = \mathbb{E}\Big[\frac{1}{b}\sum_{j=1}^b \varphi_{t,i_j}^k\Big] = \nabla F(w_t^k) - \nabla F(\tilde{w}_k). \tag{8}$$

Therefore,

$$\begin{aligned}
\mathbb{E}[\|v_t^k - \nabla F(w_t^k)\|_2^2] &= \mathbb{E}[\|\varphi_t^k + \nabla F(\tilde{w}_k) - \nabla F(w_t^k)\|_2^2] \\
&= \mathbb{E}[\|\varphi_t^k - \mathbb{E}[\varphi_t^k]\|_2^2] \\
&= \frac{1}{b^2}\mathbb{E}\left[\left\|\sum_{j=1}^b (\varphi_{t,i_j}^k - \mathbb{E}[\varphi_{t,i_j}^k])\right\|_2^2\right] \\
&= \frac{1}{b^2}\sum_{j=1}^b \mathbb{E}[\|\varphi_{t,i_j}^k - \mathbb{E}[\varphi_{t,i_j}^k]\|_2^2] \\
&\leq \frac{1}{b^2}\sum_{j=1}^b \mathbb{E}\left[\left\|\frac{\nabla f_{i_j}(w_t^k) - \nabla f_{i_j}(\tilde{w}_k)}{q_{i_j} n}\right\|_2^2\right] \\
&\leq \frac{1}{b^2}\sum_{j=1}^b \mathbb{E}\left[\frac{L_{i_j}^2}{q_{i_j}^2 n^2}\|w_t^k - \tilde{w}_k\|_2^2\right]
\end{aligned}$$

$$\leq \frac{L_\Omega^2}{b}\|w_t^k - \tilde{w}_k\|_2^2,$$

where the second and third equalities hold due to (8), and the fourth equality follows from the fact $\mathbb{E}[\|z_1+\ldots+z_r\|_2^2] = \mathbb{E}[\|z_1\|_2^2+\ldots+\|z_r\|_2^2]$, in which $z_1,\ldots,z_r$ are independent random variables with mean 0 and the first inequality comes from $\mathbb{E}[\|z-\mathbb{E}[z]\|_2^2] \leq \mathbb{E}[\|z\|_2^2]$, the second inequality employs the L_i-smoothness of f_i, and the last inequality is due to $L_\Omega \geq L_i/(nq_i)$ for $i = 1, 2, \ldots, n$. □

Lemma 3.3. *Let $c_{t_k}^k = 0$ and $c_t^k = c_{t+1}^k(1+\beta) + u_k^{\max}L_\Omega^2/(2b)$ with $\beta = b/n$ and $m = \lfloor n/b \rfloor$ for $k = 0,\ldots,K$. Assume that $b \leq n^{2/3}$ and $0 < u_k^{\max} \leq b^{3/2}/(3L_\Omega n)$. Then, the following inequality holds*

$$\left(c_{t+1}^k\left(1+\frac{1}{\beta}\right)+\frac{L_\Omega}{2}\right)\mathbb{I}_d \preceq \frac{1}{2}U_k^{-1}. \tag{9}$$

Proof. By the definition of c_t^k, recursing on t, it is easy to obtain

$$\begin{aligned} c_t^k &= \frac{u_k^{\max}L_\Omega^2}{2b}\frac{(1+\beta)^{t_k-t}-1}{\beta} \\ &\leq \frac{L_\Omega}{6b^{1/2}}\left[\left(1+\frac{b}{n}\right)^{t_k-t}-1\right] \\ &\leq \frac{L_\Omega}{6b^{1/2}}\left[\left(1+\frac{b}{n}\right)^{\lfloor n/b\rfloor}-1\right] \\ &\leq \frac{L_\Omega(e-1)}{6b^{1/2}}, \end{aligned}$$

where the first inequality holds due to $u_k^{\max} \leq b^{3/2}/(3L_\Omega n)$ and $\beta = b/n$, and the second inequality follows from $t_k \leq m = \lfloor n/b \rfloor$. In the last inequality we use the fact that (i) $\lim_{l\to+\infty}(1+1/l)^l = e$; and (ii) $(1+1/l)^l$ is an increasing function for $l > 0$ (here e is Euler's number). It follows that

$$\begin{aligned} c_{t+1}^k(1+1/\beta)+\frac{L_\Omega}{2} &\leq \frac{L_\Omega(e-1)}{6b^{1/2}}\left(1+\frac{n}{b}\right)+\frac{L_\Omega}{2} \\ &\leq \frac{L_\Omega n(e-1)}{3b^{3/2}}+\frac{L_\Omega}{2} \\ &= \frac{3L_\Omega n}{2b^{3/2}}\left[\frac{2(e-1)}{9}+\frac{b^{3/2}}{3n}\right] \\ &\leq \frac{3L_\Omega n}{2b^{3/2}}\left(\frac{4}{9}+\frac{1}{3}\right) \leq \frac{3L_\Omega n}{2b^{3/2}} \leq \frac{1}{2u_k^{\max}}, \end{aligned}$$

where the second inequality follows from $n/b \geq 1$, the third one is due to $n \geq b^{3/2}$ and $e < 3$, and the last inequality holds because $u_k^{\max} \leq b^{3/2}/(3L_\Omega n)$. Since $0 \prec U_k \preceq u_k^{\max}\mathbb{I}_d$, we have

$$[c_{t+1}^k(1+1/\beta)+\frac{L_\Omega}{2}]\mathbb{I}_d \preceq \frac{1}{2u_k^{\max}}\mathbb{I}_d \preceq \frac{1}{2}U_k^{-1}.$$

This completes the proof. □

Since the objective function is nonconvex and nonsmooth, we can not use the optimality gap $P(w)-P(w_*)$ as for the convex case or the gradient norm $\|\nabla F(w)\|_2$ as for the smooth case [8, 16, 25] to measure the convergence procedure. A popular

alternative measure is the gradient mapping [9, 26]. Here, we define the following generalized gradient mapping:

$$\mathcal{G}_{A^{-1}}(w) = A^{-1}\Big(w - \mathrm{prox}_R^{A^{-1}}\big(w - A\nabla F(w)\big)\Big), \tag{10}$$

where A is a symmetric positive definite matrix. Clearly, $\mathcal{G}_{A^{-1}}(w)$ reduces to $\nabla F(w)$ when $R(w) \equiv 0$. From [33], we know that $\mathcal{G}_{A^{-1}}(w) = 0$ if and only if w is a solution of problem (1).

Theorem 3.4. *Suppose Assumptions 1 and 2 hold, and $b \le n^{2/3}$. Let $0 < u_k^{\max} \le b^{3/2}/(3L_\Omega n)$ and $m = \lfloor n/b \rfloor$. Then, for the output w_a of Algorithm 1, after T iterations, we have sublinear convergence in expectation*

$$\mathbb{E}\Big[\|\mathcal{G}_{U_k^{-1}}(w_a)\|_{U_k}^2\Big] \le \frac{6(P(\tilde{w}_0) - P(w_*))}{T},$$

where $T = \sum_{k=0}^{K-1} t_k$ and w_* is an optimal solution of problem (1).

Proof. Recalling that the iterates of the proximal full gradient are computed by

$$\bar{w}_{t+1}^k = \mathrm{prox}_R^{U_k^{-1}}(w_t^k - U_k\nabla F(w_t^k)), \tag{11}$$

which is not actually computed in our VM-SVRG method. Applying Lemma 3.1 to the above relation (with $w' = \bar{w}_{t+1}^k$, $w = z = w_t^k$, $A = U_k$ and $\zeta = \nabla F(w_t^k)$), we take expectation and obtain

$$\mathbb{E}[P(\bar{w}_{t+1}^k)] \le \mathbb{E}\Big[P(w_t^k) + \|\bar{w}_{t+1}^k - w_t^k\|^2_{(\frac{L_\Omega}{2}\mathbb{I}_d - U_k^{-1})}\Big]. \tag{12}$$

Notice that the iterates of VM-SVRG in the inner loop are computed by

$$w_{t+1}^k = \mathrm{prox}_R^{U_k^{-1}}(w_t^k - U_k v_t^k). \tag{13}$$

Again applying Lemma 3.1 to (13) (with $w' = w_{t+1}^k$, $z = \bar{w}_{t+1}^k$, $w = w_t^k$, $A = U_k$ and $\zeta = v_t^k$), we take expectation and have

$$\begin{aligned}
&\mathbb{E}[P(w_{t+1}^k)] \\
&\le \mathbb{E}\Big[P(\bar{w}_{t+1}^k) + \frac{1}{2}\|\bar{w}_{t+1}^k - w_t^k\|^2_{(L_\Omega\mathbb{I}_d + U_k^{-1})} + \frac{1}{2}\|w_{t+1}^k - w_t^k\|^2_{(L_\Omega\mathbb{I}_d - U_k^{-1})} \\
&\quad - \frac{1}{2}\|w_{t+1}^k - \bar{w}_{t+1}^k\|^2_{U_k^{-1}} + (w_{t+1}^k - \bar{w}_{t+1}^k)^T(\nabla F(w_t^k) - v_t^k)\Big].
\end{aligned} \tag{14}$$

Summing up (12) and (14) yields

$$\begin{aligned}
\mathbb{E}[P(w_{t+1}^k)] \le \mathbb{E}\Big[&P(w_t^k) + \|\bar{w}_{t+1}^k - w_t^k\|^2_{(L_\Omega\mathbb{I}_d - \frac{1}{2}U_k^{-1})} + \frac{1}{2}\|w_{t+1}^k - w_t^k\|^2_{(L_\Omega\mathbb{I}_d - U_k^{-1})} \\
&- \frac{1}{2}\|w_{t+1}^k - \bar{w}_{t+1}^k\|^2_{U_k^{-1}} + (w_{t+1}^k - \bar{w}_{t+1}^k)^T(\nabla F(w_t^k) - v_t^k)\Big].
\end{aligned} \tag{15}$$

Let $\Gamma_t^k = (w_{t+1}^k - \bar{w}_{t+1}^k)^T(\nabla F(w_t^k) - v_t^k)$. Then the expectation on Γ_t^k can be bounded above by

$$\begin{aligned}
\mathbb{E}[\Gamma_t^k] &\le \frac{1}{2}\mathbb{E}\Big[\|w_{t+1}^k - \bar{w}_{t+1}^k\|^2_{U_k^{-1}}\Big] + \frac{1}{2}\mathbb{E}\Big[\|\nabla F(w_t^k) - v_t^k\|^2_{U_k}\Big] \\
&\le \frac{1}{2}\mathbb{E}\Big[\|w_{t+1}^k - \bar{w}_{t+1}^k\|^2_{U_k^{-1}}\Big] + \frac{u_k^{\max}L_\Omega^2}{2b}\mathbb{E}\Big[\|w_t^k - \tilde{w}_k\|_2^2\Big],
\end{aligned} \tag{16}$$

where the first inequality follows from the Cauchy-Schwarz and the Young's inequalities, and the second inequality uses the definition of $u_k^{\max}$ and Lemma 3.2. Substituting (16) into (15) yields that

$$\mathbb{E}[P(w_{t+1}^k)] \leq \mathbb{E}\Big[P(w_t^k) + \|\bar{w}_{t+1}^k - w_t^k\|^2_{(L_\Omega \mathbb{I}_d - \frac{1}{2}U_k^{-1})} + \frac{1}{2}\|w_{t+1}^k - w_t^k\|^2_{(L_\Omega \mathbb{I}_d - U_k^{-1})} + \frac{u_k^{\max} L_\Omega^2}{2b}\|w_t^k - \tilde{w}_k\|_2^2\Big]. \tag{17}$$

In order to further analyze (17), we set up the following auxiliary function:

$$\Upsilon_t^k = \mathbb{E}[P(w_t^k) + c_t^k \|w_t^k - \tilde{w}_k\|_2^2],$$

where c_t^k is defined in Lemma 3.3. Then Υ_{t+1}^k can be bounded by

$$\begin{aligned}
&\Upsilon_{t+1}^k \\
&= \mathbb{E}[P(w_{t+1}^k) + c_{t+1}^k \|w_{t+1}^k - \tilde{w}_k\|_2^2] \\
&= \mathbb{E}[P(w_{t+1}^k) + c_{t+1}^k \|w_{t+1}^k - w_t^k + w_t^k - \tilde{w}_k\|_2^2] \\
&= \mathbb{E}[P(w_{t+1}^k) + c_{t+1}^k (\|w_{t+1}^k - w_t^k\|_2^2 + \|w_t^k - \tilde{w}_k\|_2^2 + 2(w_{t+1}^k - w_t^k)^T (w_t^k - \tilde{w}_k))] \\
&\leq \mathbb{E}[P(w_{t+1}^k) + c_{t+1}^k (1 + 1/\beta)\|w_{t+1}^k - w_t^k\|_2^2 + c_{t+1}^k (1+\beta)\|w_t^k - \tilde{w}_k\|_2^2] \\
&\leq \mathbb{E}\Bigg[P(w_t^k) + \|\bar{w}_{t+1}^k - w_t^k\|^2_{(L_\Omega \mathbb{I}_d - \frac{1}{2}U_k^{-1})} + \|w_{t+1}^k - w_t^k\|^2_{(c_{t+1}^k (1+1/\beta)\mathbb{I}_d + \frac{L_\Omega}{2}\mathbb{I}_d - \frac{1}{2}U_k^{-1})} \\
&\quad + \left(c_{t+1}^k (1+\beta) + \frac{u_k^{\max} L_\Omega^2}{2b}\right)\|w_t^k - \tilde{w}_k\|_2^2\Bigg] \\
&\leq \mathbb{E}\left[P(w_t^k) + \|\bar{w}_{t+1}^k - w_t^k\|^2_{(L_\Omega \mathbb{I}_d - \frac{1}{2}U_k^{-1})} + \left(c_{t+1}^k (1+\beta) + \frac{u_k^{\max} L_\Omega^2}{2b}\right)\|w_t^k - \tilde{w}_k\|_2^2\right] \\
&= \Upsilon_t^k + \mathbb{E}\Big[\|\bar{w}_{t+1}^k - w_t^k\|^2_{(L_\Omega \mathbb{I}_d - \frac{1}{2}U_k^{-1})}\Big],
\end{aligned} \tag{18}$$

where in the first inequality we use the Cauchy-Schwarz and the Young's inequalities, the second inequality follows from (17), and the last inequality holds due to (9).

Summing (18) over $t = 0, 1, \ldots, t_k - 1$ yields that

$$\Upsilon_{t_k}^k \leq \Upsilon_0^k + \sum_{t=0}^{t_k - 1} \mathbb{E}\Big[\|\bar{w}_{t+1}^k - w_t^k\|^2_{(L_\Omega \mathbb{I}_d - \frac{1}{2}U_k^{-1})}\Big]. \tag{19}$$

The facts $c_{t_k}^k = 0$ and $\tilde{w}_{k+1} = w_{t_k}^k$ indicate that

$$\Upsilon_{t_k}^k = \mathbb{E}[P(w_{t_k}^k)] = \mathbb{E}[P(\tilde{w}_{k+1})].$$

Note that $\Upsilon_0^k = \mathbb{E}[P(w_0^k)] = \mathbb{E}[P(\tilde{w}_k)]$ holds by the fact $w_0^k = \tilde{w}_k$. It follows from (19) that

$$\mathbb{E}[P(\tilde{w}_{k+1})] \leq \mathbb{E}[P(\tilde{w}_k)] + \sum_{t=0}^{t_k - 1} \mathbb{E}\Big[\|\bar{w}_{t+1}^k - w_t^k\|^2_{(L_\Omega \mathbb{I}_d - \frac{1}{2}U_k^{-1})}\Big]. \tag{20}$$

Summing (20) over $k = 0, \ldots, K-1$ and rearranging terms, it is easy to obtain

$$\sum_{k=0}^{K-1}\sum_{t=0}^{t_k - 1} \mathbb{E}\Big[\|\bar{w}_{t+1}^k - w_t^k\|^2_{(\frac{1}{2}U_k^{-1} - L_\Omega \mathbb{I}_d)}\Big] \leq P(\tilde{w}_0) - P(\tilde{w}_K) \leq P(\tilde{w}_0) - P(w_*), \tag{21}$$

where the second inequality holds since $P(\tilde{w}_k) \geq P(w_*)$ for all $k \in \{0, 1, \ldots, K\}$.

Applying (10) with $A = U_k$ and $w = w_t^k$, and by (11), we have

$$\mathcal{G}_{U_k^{-1}}(w_t^k) = U_k^{-1}\Big(w_t^k - \mathrm{prox}_R^{U_k^{-1}}\big(w_t^k - U_k\nabla F(w_t^k)\big)\Big) = U_k^{-1}\Big(w_t^k - \bar{w}_{t+1}^k\Big).$$

Since $0 < u_k^{\max} \le b^{3/2}/(3L_\Omega n)$ and $b \le n^{2/3}$, it follows that

$$0 \prec U_k \preceq u_k^{\max}\mathbb{I}_d \preceq 1/(3L_\Omega)\mathbb{I}_d.$$

Therefore,

$$\begin{aligned}
\|\bar{w}_{t+1}^k - w_t^k\|^2_{(\frac{1}{2}U_k^{-1} - L_\Omega\mathbb{I}_d)} &= \|U_k\mathcal{G}_{U_k^{-1}}(w_t^k)\|^2_{(\frac{1}{2}U_k^{-1} - L_\Omega\mathbb{I}_d)} \\
&= \mathcal{G}_{U_k^{-1}}(w_t^k)^T U_k^T(\frac{1}{2}U_k^{-1} - L_\Omega\mathbb{I}_d)U_k\mathcal{G}_{U_k^{-1}}(w_t^k) \\
&\ge \mathcal{G}_{U_k^{-1}}(w_t^k)^T U_k^T(\frac{1}{6}U_k^{-1})U_k\mathcal{G}_{U_k^{-1}}(w_t^k) \\
&= \frac{1}{6}\big\|\mathcal{G}_{U_k^{-1}}(w_t^k)\big\|^2_{U_k}.
\end{aligned}$$

Combining the above inequality with (21) yields that

$$\sum_{k=0}^{K-1}\sum_{t=0}^{t_k-1}\frac{1}{6}\mathbb{E}\Big[\big\|\mathcal{G}_{U_k^{-1}}(w_t^k)\big\|^2_{U_k}\Big] \le P(\tilde{w}_0) - P(w_*).$$

Since the output w_a is uniformly chosen from $\{\{w_t^k\}_{t=0}^{t_k-1}\}_{k=0}^{K-1}$ and $T = \sum_{k=0}^{K-1} t_k$, we obtain

$$\mathbb{E}\Big[\|\mathcal{G}_{U_k^{-1}}(w_a)\|^2_{U_k}\Big] = \frac{1}{T}\sum_{k=0}^{K-1}\sum_{t=0}^{t_k-1}\big\|\mathcal{G}_{U_k^{-1}}(w_t^k)\big\|^2_{U_k} \le \frac{6(P(\tilde{w}_0) - P(w_*))}{T},$$

which completes the proof. □

3.2. **Linear convergence under proximal Polyak-Łojasiewicz inequality.** To achieve the desired linear convergence, we assume that $P(w)$ is a nonconvex function satisfying the proximal Polyak-Łojasiewicz (proximal-PL) inequality [13, 31], i.e., there exits a constant $\gamma > 0$ such that

$$\frac{1}{2}\mathcal{D}_R(w, \nabla F(w), \mathbb{I}_d, L) \ge \gamma(P(w) - P(w_*)), \tag{22}$$

where $\mathcal{D}_R(w, g, B, \alpha)$ is given by

$$\mathcal{D}_R(w, g, B, \alpha) = -2\alpha\min_{y\in\mathbb{R}^d}\Big\{g^T(y - w) + \frac{\alpha}{2}\|y - w\|_B^2 + R(y) - R(w)\Big\}$$

with $\alpha > 0$, $g \in \mathbb{R}^d$, and $B \in \mathbb{S}_{++}^{d\times d}$.

It has been shown that the operator $\mathcal{D}_R(w, g, B, \alpha)$ is nondecreasing in α for fixed w, g and B, see [31] for example. Here we recall the monotonic result in the following lemma.

Lemma 3.5. *(Lemma 2.3 [31]) For differentiable function F and convex function R, we have*

$$\mathcal{D}_R(w, g, B, \delta_2) \ge \mathcal{D}_R(w, g, B, \delta_1), \quad \forall\ \delta_2 \ge \delta_1 > 0,$$

where w, g and $B \in \mathbb{S}_{++}^{d\times d}$ are fixed.

Our linearly convergent method PL-VM-SVRG is presented in Algorithm 2, where VM-SVRG is employed as a subroutine.

Using the same arguments as the one in Lemma 3.3, we get the next lemma.

Algorithm 2 PL-VM-SVRG(w^0, m, b, U_0)

Input: Number of inner iterations m, initial point $\tilde{w}_0 = w^0 \in \mathbb{R}^d$, initial metric U_0, mini-batch size $b \in \{1, 2, \ldots, n\}$;

1: **for** $s = 0, 1, \ldots, S-1$ **do**
2: $\quad w^{s+1} = \text{VM-SVRG}(w^s, m, b, U_0)$.
3: **end for**

Output: w^S.

Lemma 3.6. *Let $\tilde{c}_m^k = 0$ and $\tilde{c}_t^k = \tilde{c}_{t+1}^k(1+\beta) + L_\Omega^2/(2b\theta)$ with $m = \lfloor n/b \rfloor$, $\beta = b/n$ and $\theta = L_\Omega n/b^{3/2}$ for $k = 0, \ldots, K$. Assume that $0 < u_k^{\max} \leq b^{3/2}/(6L_\Omega n)$ and $b \leq n^{2/3}$. Then, the following inequality holds*

$$\left(\tilde{c}_{t+1}^k\left(1 + \frac{1}{\beta}\right) + \frac{\theta}{2} + \frac{L_\Omega}{2}\right)\mathbb{I}_d \preceq \frac{1}{2}U_k^{-1}. \tag{23}$$

Theorem 3.7. *Suppose Assumptions 1 and 2 hold, and $0 < u_k^{\max} \leq b^{3/2}/(6L_\Omega n)$. Let $m = \lfloor n/b \rfloor$, $b \leq n^{2/3}$, $\beta = b/n$, $\theta = L_\Omega n/b^{3/2}$ and $T = Km$. Further assume the proximal-PL inequality (22) holds with the parameter $\gamma > 0$. Then, we have*

$$\mathbb{E}\left[\mathcal{D}_R\left(w_a, \nabla F(w_a), \mathbb{I}_d, \frac{1}{\underline{\alpha}}\right)\right] \leq \frac{2}{\underline{\alpha}T}\mathbb{E}[P(w^0) - P(w_*)].$$

Proof. It follows from the L-smoothness of $F(w)$ and the fact $L \leq L_\Omega$ that

$$\begin{aligned}
&F(w_{t+1}^k) \\
&\leq F(w_t^k) + \nabla F(w_t^k)^T(w_{t+1}^k - w_t^k) + \frac{L_\Omega}{2}\|w_{t+1}^k - w_t^k\|_2^2 \\
&= F(w_t^k) + (v_t^k)^T(w_{t+1}^k - w_t^k) + \frac{L_\Omega}{2}\|w_{t+1}^k - w_t^k\|_2^2 + R(w_{t+1}^k) - R(w_t^k) \\
&\quad + (\nabla F(w_t^k) - v_t^k)^T(w_{t+1}^k - w_t^k) + R(w_t^k) - R(w_{t+1}^k) \\
&= F(w_t^k) + (v_t^k)^T(w_{t+1}^k - w_t^k) + \frac{1}{2}\|w_{t+1}^k - w_t^k\|_{U_k^{-1}}^2 + R(w_{t+1}^k) - R(w_t^k) \\
&\quad + \frac{L_\Omega}{2}\|w_{t+1}^k - w_t^k\|_2^2 - \frac{1}{2}\|w_{t+1}^k - w_t^k\|_{U_k^{-1}}^2 \\
&\quad + (\nabla F(w_t^k) - v_t^k)^T(w_{t+1}^k - w_t^k) + R(w_t^k) - R(w_{t+1}^k) \\
&= F(w_t^k) + \min_{y \in \mathbb{R}^d}\left\{(v_t^k)^T(y - w_t^k) + \frac{1}{2}\|y - w_t^k\|_{U_k^{-1}}^2 + R(y) - R(w_t^k)\right\} \\
&\quad + \frac{L_\Omega}{2}\|w_{t+1}^k - w_t^k\|_2^2 - \frac{1}{2}\|w_{t+1}^k - w_t^k\|_{U_k^{-1}}^2 \\
&\quad + (\nabla F(w_t^k) - v_t^k)^T(w_{t+1}^k - w_t^k) + R(w_t^k) - R(w_{t+1}^k),
\end{aligned}$$

where the last equality follows from the definition of w_{t+1}^k in VM-SVRG.

By shifting the term $R(w_{t+1}^k)$ to the left side and using the definition of $P(w)$, we have

$$\begin{aligned}
P(w_{t+1}^k) &\leq P(w_t^k) + \min_{y \in \mathbb{R}^d}\left\{(v_t^k)^T(y - w_t^k) + \frac{1}{2}\|y - w_t^k\|_{U_k^{-1}}^2 + R(y) - R(w_t^k)\right\} \\
&\quad + \frac{L_\Omega}{2}\|w_{t+1}^k - w_t^k\|_2^2 - \frac{1}{2}\|w_{t+1}^k - w_t^k\|_{U_k^{-1}}^2 + (\nabla F(w_t^k) - v_t^k)^T(w_{t+1}^k - w_t^k) \\
&\leq P(w_t^k) + \min_{y \in \mathbb{R}^d}\left\{(v_t^k)^T(y - w_t^k) + \frac{1}{2}\|y - w_t^k\|_{U_k^{-1}}^2 + R(y) - R(w_t^k)\right\}
\end{aligned}$$

$$
\begin{aligned}
&+\frac{L_\Omega}{2}\|w_{t+1}^k-w_t^k\|_2^2-\frac{1}{2}\|w_{t+1}^k-w_t^k\|_{U_k^{-1}}^2\\
&+\frac{1}{2\theta}\|\nabla F(w_t^k)-v_t^k\|_2^2+\frac{\theta}{2}\|w_{t+1}^k-w_t^k\|_2^2\\
&=P(w_t^k)+\min_{y\in\mathbb{R}^d}\left\{(v_t^k)^T(y-w_t^k)+\frac{1}{2}\|y-w_t^k\|_{U_k^{-1}}^2+R(y)-R(w_t^k)\right\}\\
&+\left(\frac{L_\Omega}{2}+\frac{\theta}{2}\right)\|w_{t+1}^k-w_t^k\|_2^2-\frac{1}{2}\|w_{t+1}^k-w_t^k\|_{U_k^{-1}}^2+\frac{1}{2\theta}\|\nabla F(w_t^k)-v_t^k\|_2^2,
\end{aligned}
\tag{24}
$$

where the second inequality follows from the Cauchy-Schwarz and the Young's inequalities. By noting $\mathbb{E}[v_t^k]=\nabla F(w_t^k)$ and using the definition of $\mathcal{D}_R$, we take expectation on both sides of (24) conditioned on w_t^k and obtain

$$
\begin{aligned}
\mathbb{E}[P(w_{t+1}^k)]&\le\mathbb{E}[P(w_t^k)]-\frac{1}{2}\mathcal{D}_R(w_t^k,\nabla F(w_t^k),U_k^{-1},1)+\left(\frac{\theta}{2}+\frac{L_\Omega}{2}\right)\mathbb{E}[\|w_{t+1}^k-w_t^k\|_2^2]\\
&\quad-\frac{1}{2}\mathbb{E}[\|w_{t+1}^k-w_t^k\|_{U_k^{-1}}^2]+\frac{1}{2\theta}\mathbb{E}[\|\nabla F(w_t^k)-v_t^k\|_2^2]\\
&\le\mathbb{E}[P(w_t^k)]-\frac{1}{2}\mathcal{D}_R(w_t^k,\nabla F(w_t^k),U_k^{-1},1)+\left(\frac{\theta}{2}+\frac{L_\Omega}{2}\right)\mathbb{E}[\|w_{t+1}^k-w_t^k\|_2^2]\\
&\quad-\frac{1}{2}\mathbb{E}[\|w_{t+1}^k-w_t^k\|_{U_k^{-1}}^2]+\frac{L_\Omega^2}{2b\theta}\|w_t^k-\tilde{w}_k\|_2^2,
\end{aligned}
\tag{25}
$$

where the last inequality follows from Lemma 3.2.

In order to analyze the convergence rate of PL-VM-SVRG, we consider the following auxiliary function

$$
\Psi_t^k=\mathbb{E}[P(w_t^k)+\tilde{c}_t^k\|w_t^k-\tilde{w}_k\|_2^2],
$$

where $\tilde{c}_t^k$ is defined in Lemma 3.6. Then we can derive an upper bound on Ψ_{t+1}^k as follows.

$$
\begin{aligned}
&\Psi_{t+1}^k\\
&=\mathbb{E}[P(w_{t+1}^k)+\tilde{c}_{t+1}^k\|w_{t+1}^k-\tilde{w}_k\|_2^2]\\
&=\mathbb{E}[P(w_{t+1}^k)+\tilde{c}_{t+1}^k(\|w_{t+1}^k-w_t^k\|_2^2+\|w_t^k-\tilde{w}_k\|_2^2+2(w_{t+1}^k-w_t^k)^T(w_t^k-\tilde{w}_k))]\\
&\le\mathbb{E}[P(w_{t+1}^k)+\tilde{c}_{t+1}^k(1+1/\beta)\|w_{t+1}^k-w_t^k\|_2^2+\tilde{c}_{t+1}^k(1+\beta)\|w_t^k-\tilde{w}_k\|_2^2]\\
&\le\mathbb{E}\left[P(w_t^k)+\left(\tilde{c}_{t+1}^k(1+\beta)+\frac{L_\Omega^2}{2b\theta}\right)\|w_t^k-\tilde{w}_k\|_2^2\right]-\frac{1}{2}\mathcal{D}_R(w_t^k,\nabla F(w_t^k),U_k^{-1},1)\\
&\quad+\left(\frac{\theta}{2}+\frac{L_\Omega}{2}+\tilde{c}_{t+1}^k(1+1/\beta)\right)\mathbb{E}[\|w_{t+1}^k-w_t^k\|_2^2]-\frac{1}{2}\mathbb{E}[\|w_{t+1}^k-w_t^k\|_{U_k^{-1}}^2]\\
&=\Psi_t^k-\frac{1}{2}\mathcal{D}_R(w_t^k,\nabla F(w_t^k),U_k^{-1},1)\\
&\quad+\left(\frac{\theta}{2}+\frac{L_\Omega}{2}+\tilde{c}_{t+1}^k(1+1/\beta)\right)\mathbb{E}[\|w_{t+1}^k-w_t^k\|_2^2]-\frac{1}{2}\mathbb{E}[\|w_{t+1}^k-w_t^k\|_{U_k^{-1}}^2]\\
&\le\Psi_t^k-\frac{1}{2}\mathcal{D}_R(w_t^k,\nabla F(w_t^k),U_k^{-1},1),
\end{aligned}
\tag{26}
$$

where the first inequality is due to the Cauchy-Schwarz and the Young's inequalities, the second inequality follows from (25), the last inequality holds because the

sequence of $\tilde{c}_t^k$ satisfies (23) while the last equality uses the definition of Ψ_t^k. Rearranging terms of (26) yields

$$\mathcal{D}_R(w_t^k, \nabla F(w_t^k), U_k^{-1}, 1) \leq 2(\Psi_t^k - \Psi_{t+1}^k). \tag{27}$$

Summing (27) over $t = 0, 1, \ldots, m-1$, we get

$$\sum_{t=0}^{m-1} \mathcal{D}_R(w_t^k, \nabla F(w_t^k), U_k^{-1}, 1) \leq 2(\Psi_0^k - \Psi_m^k). \tag{28}$$

Since $\tilde{c}_m^k = 0$, from the definition of $\tilde{w}_{k+1}$ in VM-SVRG, we have $\Psi_m^k = \mathbb{E}[P(w_m^k)] = \mathbb{E}[P(\tilde{w}_{k+1})]$. Recall that $w_0^k = \tilde{w}_k$, we get $\Psi_0^k = \mathbb{E}[P(w_0^k)] = \mathbb{E}[P(\tilde{w}_k)]$. Therefore, it follows from (28) that

$$\sum_{t=0}^{m-1} \mathcal{D}_R(w_t^k, \nabla F(w_t^k), U_k^{-1}, 1) \leq 2\mathbb{E}[P(\tilde{w}_k) - P(\tilde{w}_{k+1})]. \tag{29}$$

Summing up (29) for $k = 0, 1, \ldots, K-1$, and multiplying both sides with $\frac{1}{T}$, we have

$$\frac{1}{T} \sum_{k=0}^{K-1} \sum_{t=0}^{m-1} \mathcal{D}_R(w_t^k, \nabla F(w_t^k), U_k^{-1}, 1) \leq \frac{2}{T}\mathbb{E}[P(\tilde{w}_0) - P(\tilde{w}_K)]. \tag{30}$$

By $u_k^{\min}\mathbb{I}_d \preceq U_k \preceq u_k^{\max}\mathbb{I}_d$, we have $\|y - w_t^k\|_{U_k^{-1}}^2 \leq \frac{1}{u_k^{\min}}\|y - w_t^k\|_2^2$ for any $y \in \mathbb{R}^d$. Then,

$$\begin{aligned}
&\nabla F(w_t^k)^T(y - w_t^k) + \frac{1}{2}\|y - w_t^k\|_{U_k^{-1}}^2 + R(y) - R(w_t^k) \\
\leq\ &\nabla F(w_t^k)^T(y - w_t^k) + \frac{1}{2u_k^{\min}}\|y - w_t^k\|_2^2 + R(y) - R(w_t^k) \\
\leq\ &\nabla F(w_t^k)^T(y - w_t^k) + \frac{1}{2\underline{\alpha}}\|y - w_t^k\|_2^2 + R(y) - R(w_t^k),
\end{aligned}$$

where the last inequality is due to $u_k^{\min} \geq \underline{\alpha}$. Note that if $f_1(y) \leq f_2(y)$ for all y, then $\min_y f_1(y) \leq \min_y f_2(y)$. Consequently,

$$\begin{aligned}
&\min_{y\in\mathbb{R}^d} \left\{\nabla F(w_t^k)^T(y - w_t^k) + \frac{1}{2}\|y - w_t^k\|_{U_k^{-1}}^2 + R(y) - R(w_t^k)\right\} \\
\leq &\min_{y\in\mathbb{R}^d} \left\{\nabla F(w_t^k)^T(y - w_t^k) + \frac{1}{2\underline{\alpha}}\|y - w_t^k\|_2^2 + R(y) - R(w_t^k)\right\}.
\end{aligned}$$

It follows from the definition of $\mathcal{D}_R(w_t^k, \nabla F(w_t^k), U_k^{-1}, 1)$ that

$$\begin{aligned}
&\mathcal{D}_R(w_t^k, \nabla F(w_t^k), U_k^{-1}, 1) \\
\geq\ &-2\min_{y\in\mathbb{R}^d} \left\{\nabla F(w_t^k)^T(y - w_t^k) + \frac{1}{2\underline{\alpha}}\|y - w_t^k\|_2^2 + R(y) - R(w_t^k)\right\} \\
=\ &\underline{\alpha}\mathcal{D}_R\left(w_t^k, \nabla F(w_t^k), \mathbb{I}_d, \frac{1}{\underline{\alpha}}\right).
\end{aligned} \tag{31}$$

Combining (30) and (31) yields that

$$\begin{aligned}
&\frac{\underline{\alpha}}{T} \sum_{k=0}^{K-1} \sum_{t=0}^{m-1} \mathcal{D}_R\left(w_t^k, \nabla F(w_t^k), \mathbb{I}_d, \frac{1}{\underline{\alpha}}\right) \\
\leq\ &\frac{1}{T} \sum_{k=0}^{K-1} \sum_{t=0}^{m-1} \mathcal{D}_R(w_t^k, \nabla F(w_t^k), U_k^{-1}, 1)
\end{aligned}$$

$$\leq \frac{2}{T}\mathbb{E}[P(\tilde{w}_0) - P(\tilde{w}_K)].$$

Since the output w_a of VM-SVRG is uniformly chose from $\{\{w_t^k\}_{t=0}^{m-1}\}_{k=0}^{K-1}$, we have

$$\begin{aligned}\mathbb{E}\left[\mathcal{D}_R\left(w_a, \nabla F(w_a), \mathbb{I}_d, \frac{1}{\underline{\alpha}}\right)\right] &= \frac{1}{T}\sum_{k=0}^{K-1}\sum_{t=0}^{m-1}\mathcal{D}_R\left(w_t^k, \nabla F(w_t^k), \mathbb{I}_d, \frac{1}{\underline{\alpha}}\right)\\ &\leq \frac{2}{\underline{\alpha}T}\mathbb{E}[P(\tilde{w}_0) - P(\tilde{w}_K)]\\ &\leq \frac{2}{\underline{\alpha}T}\mathbb{E}[P(w^0) - P(w_*)], \end{aligned} \tag{32}$$

where the last inequality follows from the fact that $P(w) \geq P(w_*)$ for any $w \in \mathbb{R}^d$ and $w^0 = \tilde{w}_0$. □

Notice that both Theorems 3.4 and 3.7 show the sublinear convergence of VM-SVRG. However, Theorem 3.7 employs a different measure which is useful in establishing the linear convergence of PL-VM-SVRG.

Theorem 3.8. *Under the same conditions as Theorem 3.7, and set* $T = \lceil 2/(\gamma\underline{\alpha})\rceil$, *then we have linear convergence in expectation*

$$\mathbb{E}[P(w^S) - P(w_*)] \leq (2^{-S})(P(w^0) - P(w_*)).$$

Proof. Recalling that in each iteration of PL-VM-SVRG w^s is the input of VM-SVRG while w^{s+1} is the output. By replacing w^0 and w_a in (32) with w^s and w^{s+1}, respectively, we obtain

$$\mathbb{E}\left[\mathcal{D}_R\left(w^{s+1}, \nabla F(w^{s+1}), \mathbb{I}_d, \frac{1}{\underline{\alpha}}\right)\right] \leq \frac{2}{\underline{\alpha}T}\mathbb{E}[P(w^s) - P(w_*)]. \tag{33}$$

Since $u_k^{\max} \leq b^{3/2}/(6L_\Omega n)$ and $n \geq b^{3/2}$, we have $u_k^{\max} \leq 1/(6L_\Omega) < 1/L_\Omega$, which together with $\underline{\alpha} \leq u_k^{\min} \leq u_k^{\max} \leq 1/L_\Omega$ implies that $L_\Omega \leq 1/\underline{\alpha}$. It follows from Lemma 3.5 and $L \leq L_\Omega$ that

$$\begin{aligned}\mathcal{D}_R(w_t^k, \nabla F(w_t^k), \mathbb{I}_d, L) &\leq \mathcal{D}_R(w_t^k, \nabla F(w_t^k), \mathbb{I}_d, L_\Omega)\\ &\leq \mathcal{D}_R\left(w_t^k, \nabla F(w_t^k), \mathbb{I}_d, \frac{1}{\underline{\alpha}}\right).\end{aligned} \tag{34}$$

Using the proximal-PL inequality with $w = w^{s+1}$ and taking expectation, we have

$$2\gamma\mathbb{E}[P(w^{s+1}) - P(w_*)] \leq \mathbb{E}[\mathcal{D}_R(w^{s+1}, \nabla F(w^{s+1}), \mathbb{I}_d, L)]. \tag{35}$$

Combining (33), (34) and (35), and substituting the specific value of T, we obtain

$$\begin{aligned}\mathbb{E}[P(w^{s+1}) - P(w_*)] &\leq \frac{1}{2\gamma}\frac{2}{\underline{\alpha}T}\mathbb{E}[P(w^s) - P(w_*)]\\ &= \frac{1}{\gamma\underline{\alpha}T}\mathbb{E}[P(w^s) - P(w_*)]\\ &\leq \frac{1}{2}\mathbb{E}[P(w^s) - P(w_*)]\end{aligned}$$

Applying the above inequality recursively we will get the desired result. □

3.3. **Comparisons of complexity.** In order to measure the efficiency of a proximal stochastic algorithm, we employ the stochastic first-order oracle (SFO) and proximal oracle (PO) complexity. In particular, for a given point $w \in \mathbb{R}^d$, an SFO takes an index $i \in \{1, 2, \ldots, n\}$ and returns $\nabla f_i(w)$ [9], and a PO returns an output of a proximal problem [26].

From Theorem 3.4, we conclude that VM-SVRG requires $\mathcal{O}(1/\epsilon)$ total number of inner iterations to achieve $\mathbb{E}[\|\mathcal{G}_{U_k^{-1}}(w_a)\|_{U_k}^2] \leq \epsilon$. Hence the PO complexity is $\mathcal{O}(1/\epsilon)$ as one PO is involved in each inner iteration. Let $b = n^{2/3}$. Then the SFO complexity in all inner iterations is $\mathcal{O}(n^{2/3}/\epsilon)$. Recall that VM-SVRG takes n SFO to compute the average gradient in an outer iteration and T is at most a multiple of m. By summing the total cost together, we obtain the SFO complexity of VM-SVRG is $\mathcal{O}(n + (n^{2/3}/\epsilon))$.

When the objective function satisfies the proximal-PL inequality, Theorem 3.8 indicates that, for $b = n^{2/3}$ and $T = \mathcal{O}(\kappa)$ with $\kappa = L_\Omega/\gamma$, to achieve $\mathbb{E}[P(w^S) - P(w_*)] \leq \epsilon$, the SFO and PO complexity of Algorithm 2 are $\mathcal{O}((n+\kappa n^{2/3})\log(1/\epsilon))$ and $\mathcal{O}(\kappa \log(1/\epsilon))$, respectively.

TABLE 1. Comparison of the SFO and PO complexity.

Complexity	Prox-GD	Prox-SGD	Prox-SVRG	VM-SVRG
SFO	$\mathcal{O}(n/\epsilon)$	$\mathcal{O}(1/\epsilon^2)$	$\mathcal{O}(n + (n^{2/3}/\epsilon))$	$\mathcal{O}(n + (n^{2/3}/\epsilon))$
PO	$\mathcal{O}(1/\epsilon)$	$\mathcal{O}(1/\epsilon)$	$\mathcal{O}(1/\epsilon)$	$\mathcal{O}(1/\epsilon)$
SFO(PL)	$\mathcal{O}(n\kappa \log(1/\epsilon))$	$\mathcal{O}(1/\epsilon^2)$	$\mathcal{O}((n + \kappa n^{2/3})\log(1/\epsilon))$	$\mathcal{O}((n + \kappa n^{2/3})\log(1/\epsilon))$
PO(PL)	$\mathcal{O}(\kappa \log(1/\epsilon))$	$\mathcal{O}(1/\epsilon)$	$\mathcal{O}(\kappa \log(1/\epsilon))$	$\mathcal{O}(\kappa \log(1/\epsilon))$

Table 1 lists the SFO and PO complexity of different methods for the above cases, where PL represents the proximal-PL inequality case. It is easy to see that, for each case, the SFO and PO complexity of VM-SVRG are lower than that of Prox-GD and Prox-SGD, and are the same as Prox-SVRG.

4. **Numerical experiments.** In this section, we present numerical comparisons of VM-SVRG and some recent developed proximal SVRG methods on four standard data sets listed in Table 2, which can be downloaded from the LIBSVM website [1]. For fair comparison, all methods are implemented in Matlab 2018b under Windows 10 operating system on a laptop with an Intel Core i7, 1.80 GHz processor and 16 GB of RAM.

TABLE 2. The information of data sets.

Data sets	n	d
ijcnn1	49,990	22
rcv1	20,242	47,236
real-sim	72,309	20,958
covtype	581,012	54

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/

We focus on two standard testing problems in machine learning, i.e., the elastic net regularized logistic regression (LR) problem

$$\text{LR} \quad \min_{w\in\mathbb{R}^d} P(w) = \frac{1}{n}\sum_{i=1}^{n} \log(1+\exp(-c_i a_i^T w)) + \frac{\lambda_2}{2}\|w\|_2^2 + \lambda_1\|w\|_1, \tag{36}$$

and the sparse nonconvex support vector machine (SVM) problem with a sigmoid loss function

$$\text{SVM} \quad \min_{w\in\mathbb{R}^d} P(w) = \frac{1}{n}\sum_{i=1}^{n} (1-\tanh(c_i a_i^T w)) + \lambda_1\|w\|_1, \tag{37}$$

where λ_1 and λ_2 are two nonnegative regularization parameters, and $\{(a_i, c_i)\}_{i=1}^n$ is a set of training examples with $a_i \in \mathbb{R}^d$ being the feature vector and $c_i \in \{-1, +1\}$ being the corresponding label.

For the LR model, as suggested in [34], the test was performed with $R(w) = \lambda_1\|w\|_1$ and

$$f_i(w) = \log(1+\exp(-c_i a_i^T w)) + \frac{\lambda_2}{2}\|w\|_2^2,$$

where $\lambda_1 = 10^{-5}$ and $\lambda_2 = 10^{-4}$ for ijcnn1, rcv1 and real-sim, and $\lambda_1 = 10^{-4}$ and $\lambda_2 = 10^{-5}$ for covtype. For the SVM model, as suggested in [31], $\lambda_1 = 10^{-5}$ was used for all data sets. The Lipschitz constants are set to $L_i = \|a_i\|_2^2/4 + \lambda_2$. We set $L = \max_{i=1,2,\ldots,n} L_i$. As suggested in [23], we set $\omega = 10^{-6}$.

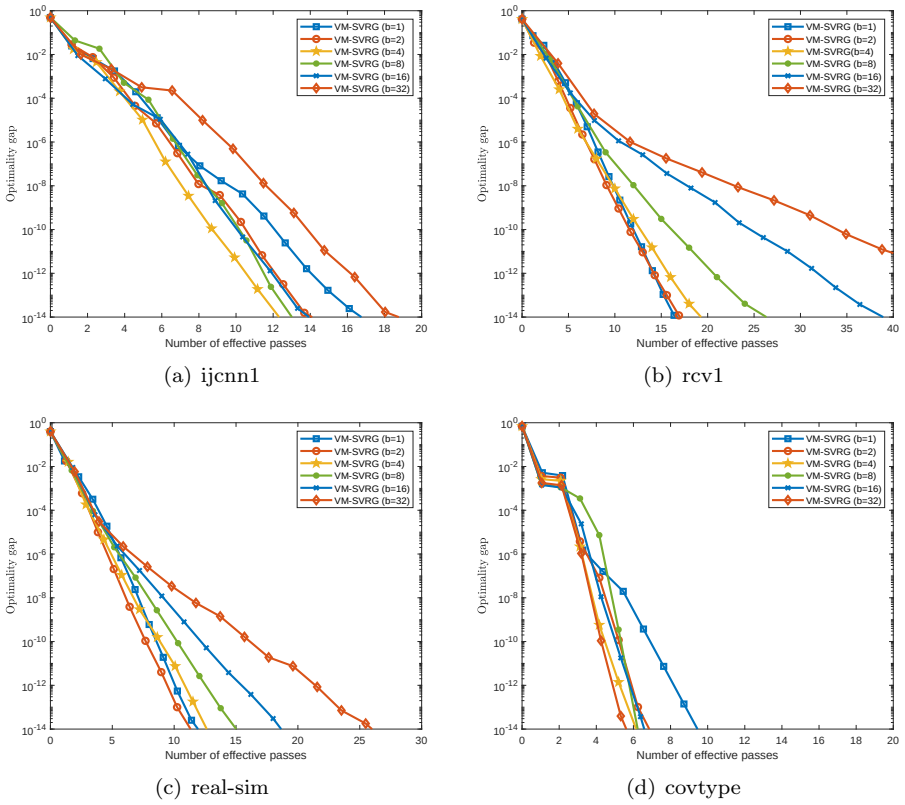
In all the following figures, the x-axis is the number of effective passes over the data set, where the evaluation of n component gradients counts as one effective pass. In experiments on LR model, the y-axis with "Optimality gap" denotes the value $P(\tilde{w}_k) - P(w_*)$ with w_* obtained by running Prox-SVRG with best-tuned fixed stepsizes. In experiments on SVM model, the y-axis is the squared norm of gradient.

4.1. **Experiment results on LR.** We first tested VM-SVRG with different values of b on the four data sets listed in Table 2 to investigate the influence of mini-batch size. Figure 1 presents the results of VM-SVRG with $b = 1, 2, 4, 8, 16, 32$. We see that when the mini-batch size increases to $b = 2, 4, 8, 16$, the performance of VM-SVRG is better than or comparable to the case $b = 1$.

Then we compared VM-SVRG with mS2GD [14] and mSARAH, which are variants of Prox-SVRG [34] and SARAH [20] in the proximal mini-batch setting, respectively. We also compared the mS2GD-BB method, which was obtained by combining mS2GD with the BB method. Moreover, a mini-batch proximal variant of SARAH-BB [18], named mSARAH-BB, was also run for comparison.

For mS2GD, mS2GD-BB, mSARAH and mSARAH-BB, we set $b = 8$ which performs better than other values in our test. In addition, m and initial stepsizes were tuned by hand to get the best performance. For our VM-SVRG method, we set $b = 4$ and used best-tuned parameters. The best choices of m for mS2GD, mS2GD-BB, mSARAH, mSARAH-BB and VM-SVRG, as well as the best-tuned stepsizes η for mS2GD and mSARAH, are presented in Table 3.

From Figure 2 we see that our VM-SVRG performs better than or comparable to other four methods.



(a) ijcnn1 (b) rcv1

(c) real-sim (d) covtype

FIGURE 1. Comparison of VM-SVRG with different mini-batch sizes.

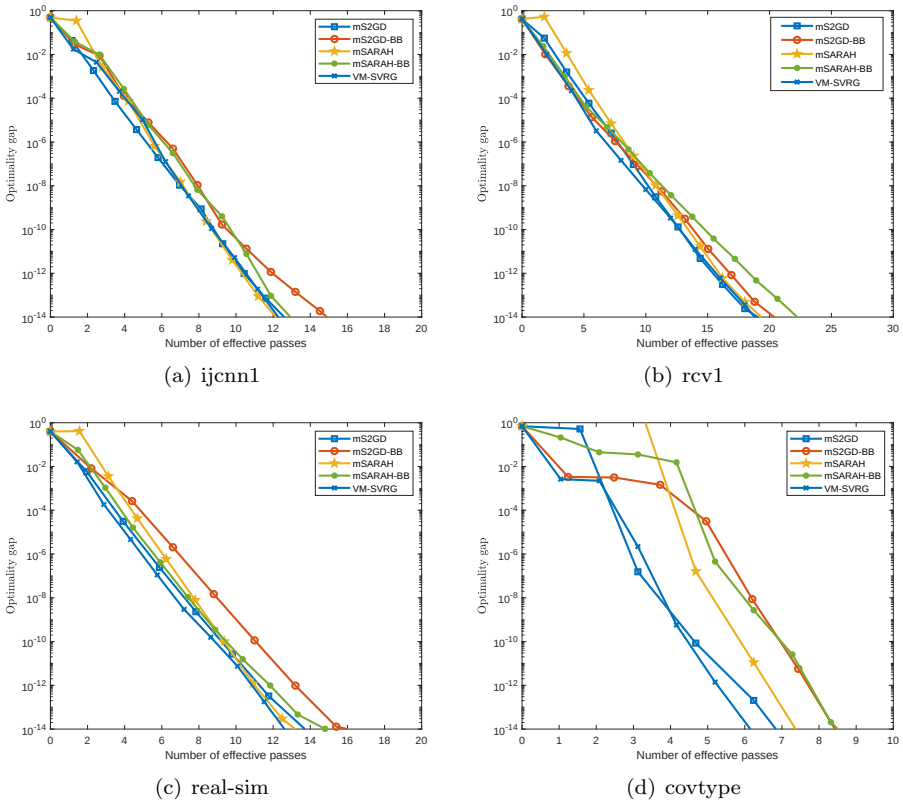
TABLE 3. Best choices of parameters for the methods.

Data sets	mS2GD(m,η)	mS2GD-BB	mSARAH(m,η)	mSARAH-BB	VM-SVRG
ijcnn1	$(0.02n, \frac{4}{L})$	$0.04n$	$(0.05n, \frac{1.8}{L})$	$0.04n$	$0.04n$
rcv1	$(0.1n, \frac{4}{L})$	$0.11n$	$(0.1n, \frac{3.5}{L})$	$0.09n$	$0.25n$
real-sim	$(0.12n, \frac{0.6}{L})$	$0.15n$	$(0.07n, \frac{2}{L})$	0.06	$0.11n$
covtype	$(0.07n, \frac{21}{L})$	$0.03n$	$(0.07n, \frac{25}{L})$	$0.008n$	$0.01n$

4.2. **Experiment results on SVM.** Now we apply VM-SVRG to the SVM model (37). Since the values of $\overline{\alpha}_k$ and $\underline{\alpha}_k$ may be extremely large in the nonconvex case, we project them into $[10^{-6}, 2/L_\Omega]$ in our test to chop those values.

We also compared VM-SVRG with different mini-batch sizes b. It can be seen from Figure 3 that, similarly to the LR model, by increasing the mini-batch size to $b = 2, 4, 8, 16$, VM-SVRG performs better than or comparable to that with $b = 1$.

Then we compared VM-SVRG with other SGD methods. Since [31] showed that Prox-SVRG performs better than Prox-GD for solving (37), we do not present the results of Prox-GD. Figure 4 presents VM-SVRG vs. mS2GD, where we set b to 4 and 8 for VM-SVRG and mS2GD, respectively, and use best-tuned values for other parameters. Clearly, VM-SVRG outperforms mS2GD in the sense of squared norm of gradient.



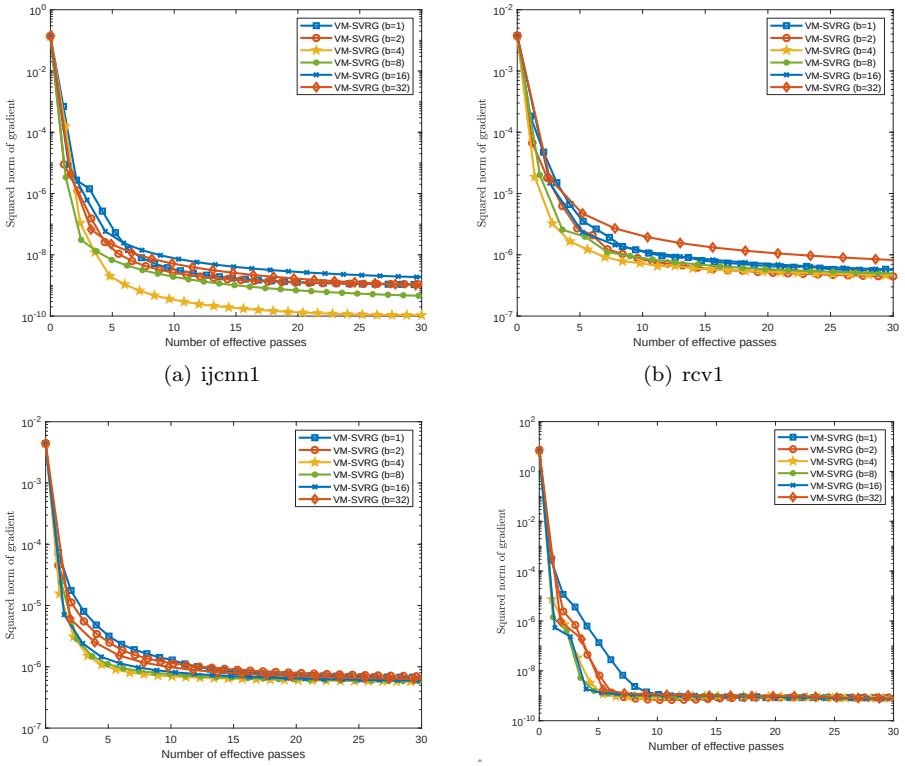
(a) ijcnn1 (b) rcv1

(c) real-sim (d) covtype

FIGURE 2. Comparison of VM-SVRG and other modern methods for solving LR problem.

5. **Conclusion.** We proposed a variable metric mini-batch proximal stochastic variance reduced gradient method VM-SVRG for nonconvex nonsmooth optimization, which uses a diagonal Barzilai-Borwein stepsize to update the metric. We showed that VM-SVRG converges sublinearly to a stationary point in expectation. Based on the proximal Polyak-Łojasiewicz inequality, the sublinear rate was further improved to linear by slightly modifying VM-SVRG. The complexity of the proposed methods was lower than that of Prox-GD and Prox-SGD, and was the same as Prox-SVRG under different conditions. Numerical results showed that our VM-SVRG method is better than or comparable to the state-of-the-art proximal stochastic gradient methods.

Acknowledgments. The research was supported in part by the National Natural Science Foundation of China (Grant Nos. 12071108, 11701137, 11671116, 11631013, 11991020, 12021001), the Major Research Plan of the National Natural Science Foundation of China (Grant No. 91630202), Beijing Academy of Artificial Intelligence (BAAI), and the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDA27000000).



(a) ijcnn1 (b) rcv1

(c) real-sim (d) covtype

FIGURE 3. Comparison of VM-SVRG with different mini-batch sizes.

REFERENCES

[1] J. Barzilai and J. M. Borwein, Two-point step size gradient methods, *IMA J. Numer. Anal.*, **8** (1988), 141–148.

[2] J. F. Bonnans, J. Ch. Gilbert, C. Lemaréchal and C. A. Sagastizábal, A family of variable metric proximal methods, *Math. Programming*, **68** (1995), 15–47.

[3] L. Bottou, F. E. Curtis and J. Nocedal, Optimization methods for large-scale machine learning, *SIAM Rev.*, **60** (2018), 223–311.

[4] Y.-H. Dai, M. Al-Baali and X. Yang, A positive Barzilai–Borwein-like stepsize and an extension for symmetric linear systems, *Numerical Analysis and Optimization*, **134** (2015), 59–75.

[5] Y.-H. Dai, Y. Huang and X.-W. Liu, A family of spectral gradient methods for optimization, *Comput. Optim. Appl.*, **74** (2019), 43–65.

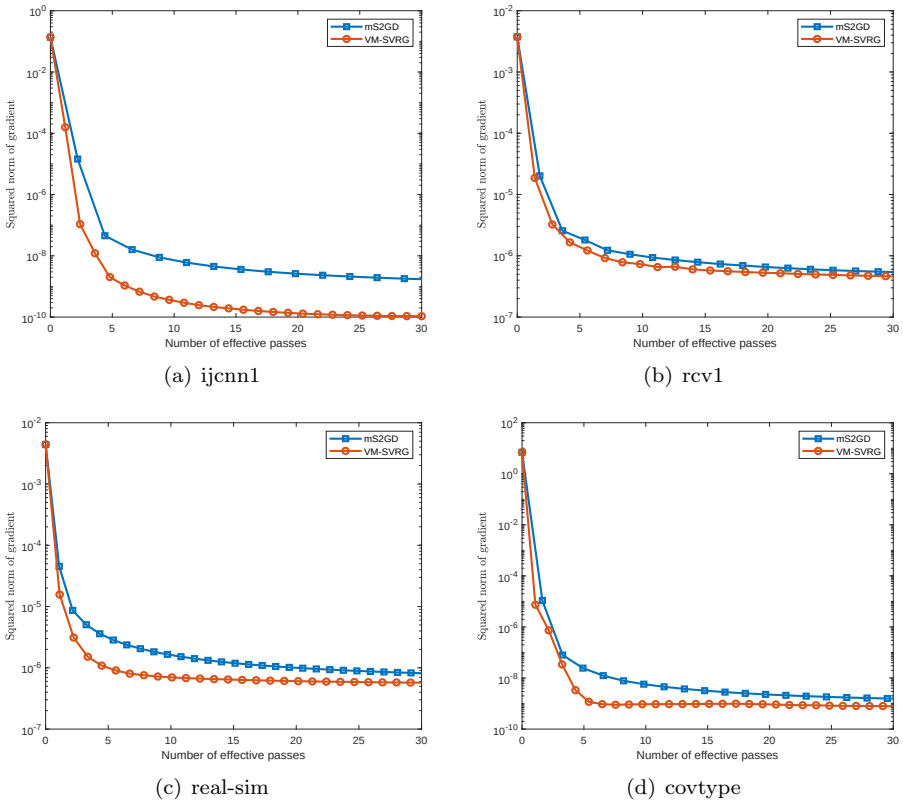
[6] A. Defazio, F. Bach and S. Lacoste-Julien, SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives, in *Advances in Neural Information Processing Systems*, (2014), 1646–1654.

[7] R. Fletcher, On the Barzilai–Borwein method, in *Optimization and control with applications*, Springer, New York, **96** (2005), 235–256.

[8] S. Ghadimi and G. Lan, Stochastic first-and zeroth-order methods for nonconvex stochastic programming, *SIAM J. Optim.*, **23** (2013), 2341–2368.

[9] S. Ghadimi, G. Lan and H. Zhang, Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization, *Math. Program.*, **155** (2016), 267–305.

[10] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics. Springer, New York, 2009.



(a) ijcnn1 (b) rcv1

(c) real-sim (d) covtype

FIGURE 4. Comparison of VM-SVRG and mS2GD for solving SVM problem.

[11] Y. Huang, Y.-H. Dai, X.-W. Liu and H. Zhang, Gradient methods exploiting spectral properties, *Optim. Methods Softw.*, **35** (2020), 681–705.

[12] R. Johnson and T. Zhang, Accelerating stochastic gradient descent using predictive variance reduction, in *Advances in Neural Information Processing Systems*, (2013), 315–323.

[13] H. Karimi, J. Nutini and M. Schmidt, Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition, in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, (2016), 795–811.

[14] J. Konečnỳ, J. Liu, P. Richtárik and M. Takáč, Mini-batch semi-stochastic gradient descent in the proximal setting, *IEEE Journal of Selected Topics in Signal Processing*, **10** (2015), 242–255.

[15] J. Konečnỳ and P. Richtárik, Semi-stochastic gradient descent methods, *Frontiers in Applied Mathematics and Statistics*, **3** (2017), 9.

[16] L. Lei, C. Ju, J. Chen and M. I. Jordan, Non-convex finite-sum optimization via SCSG methods, in *Advances in Neural Information Processing Systems*, (2017), 2348–2358.

[17] Z. Li and J. Li, A simple proximal stochastic gradient method for nonsmooth nonconvex optimization, in *Advances in Neural Information Processing Systems*, (2018), 5564–5574.

[18] Y. Liu, X. Wang and T. Guo, A linearly convergent stochastic recursive gradient method for convex optimization, *Optim. Lett.*, **14** (2020), 2265–2283.

[19] Y. Nesterov, *Introductory Lectures on Convex Programming: A Basic Course*, Applied Optimization, 87. Kluwer Academic Publishers, Boston, MA, 2004.

[20] L. M. Nguyen, J. Liu, K. Scheinberg and M. Takáč, SARAH: A novel method for machine learning problems using stochastic recursive gradient, in *Proceedings of the 34th International Conference on Machine Learning*, **70** (2017), 2613–2621.

[21] L. A. Parente, P. A. Lotito and M. V. Solodov, A class of inexact variable metric proximal point algorithms, *SIAM J. Optim.*, **19** (2008), 240–260.
[22] N. Parikh, S. Boyd et al., Proximal algorithms, *Foundations and Trends in Optimization*, **1** (2014), 127–239.
[23] Y. Park, S. Dhar, S. Boyd and M. Shah, Variable metric proximal gradient method with diagonal Barzilai-Borwein stepsize, (2019), arXiv:1910.07056.
[24] N. H. Pham, L. M. Nguyen, D. T. Phan and Q. Tran-Dinh, ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization, *J. Mach. Learn. Res.*, **21** (2020), Paper No. 110, 48 pp.
[25] S. J. Reddi, A. Hefny, S. Sra, B. Poczos and A. Smola, Stochastic variance reduction for nonconvex optimization, in *International Conference on Machine Learning*, (2016), 314–323.
[26] S. J. Reddi, S. Sra, B. Poczos and A. J. Smola, Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization, in *Advances in Neural Information Processing Systems*, (2016), 1145–1153.
[27] H. Robbins and S. Monro, A stochastic approximation method, *Ann. Math. Statistics*, **22** (1951), 400–407.
[28] M. Schmidt, N. Le Roux and F. Bach, Minimizing finite sums with the stochastic average gradient, *Math. Program.*, **162** (2017), 83–112.
[29] F. Shang, K. Zhou, H. Liu, J. Cheng, I. W. Tsang, L. Zhang, D. Tao and L. Jiao, VR-SGD: A simple stochastic variance reduction method for machine learning, *IEEE Transactions on Knowledge and Data Engineering*, **32** (2020), 188–202.
[30] C. Tan, S. Ma, Y.-H. Dai and Y. Qian, Barzilai-Borwein step size for stochastic gradient descent, in *Advances in Neural Information Processing Systems*, (2016), 685–693.
[31] X. Wang, X. Wang and Y.-X. Yuan, Stochastic proximal quasi-Newton methods for non-convex composite optimization, *Optim. Methods Softw.*, **34** (2019), 922–948.
[32] X. Wang, S. Wang and H. Zhang, Inexact proximal stochastic gradient method for convex composite optimization, *Comput. Optim. Appl.*, **68** (2017), 579–618.
[33] X. Wang and H. Zhang, Inexact proximal stochastic second-order methods for nonconvex composite optimization, *Optim. Methods Softw.*, **35** (2020), 808–835.
[34] L. Xiao and T. Zhang, A proximal stochastic gradient method with progressive variance reduction, *SIAM J. Optim.*, **24** (2014), 2057–2075.
[35] T. Yu, X.-W. Liu, Y.-H. Dai and J. Sun, A minibatch proximal stochastic recursive gradient algorithm using a trust-region-like scheme and Barzilai-Borwein stepsizes, *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
[36] T. Yu, X.-W. Liu, Y.-H. Dai and J. Sun, Stochastic variance reduced gradient methods using a trust-region-like scheme, *J. Sci. Comput.*, **87** (2021), Article number: 5.
[37] T. Yu, X.-W. Liu, Y.-H. Dai and J. Sun, A variable metric mini-batch proximal stochastic recursive gradient algorithm with diagonal Barzilai-Borwein stepsize, 2020, arXiv:2010.00817.

Received January 2021; revised February 2021.

E-mail address: yuteng206@163.com
E-mail address: mathlxw@hebut.edu.cn
E-mail address: dyh@lsec.cc.ac.cn
E-mail address: jsun@nus.edu.sg