

Improved Optimization of Finite Sums with Minibatch Stochastic Variance Reduced Proximal Iterations

Jialei Wang^{*} and Tong Zhang[#]

^{*}Department of Computer Science, University of Chicago, IL, USA

[#]Tencent AI Lab, Shenzhen, China

June 22, 2017

Abstract

We present novel minibatch stochastic optimization methods for empirical risk minimization problems, the methods efficiently leverage variance reduced first-order and sub-sampled higher-order information to accelerate the convergence speed. For quadratic objectives, we prove improved iteration complexity over state-of-the-art under reasonable assumptions. We also provide empirical evidence of the advantages of our method compared to existing approaches in the literature.

1 Introduction

We consider the following optimization problem of finite-sums:

$$\min_w f(w) := \frac{1}{n} \sum_{i=1}^n f_i(w), \quad (1)$$

which arises in many machine learning problems. Let $x_1, x_2, \dots, x_n \in \mathbb{R}^d$ be feature vectors of n data samples, and $b_1, \dots, b_n \in \mathbb{R}$ or $\{-1, 1\}$ be the corresponding target variable of interest. (1) covers many popular models used in machine learning, for example when $f_i(w) = \frac{1}{2}(b_i - \langle w, x_i \rangle)^2 + \frac{\lambda}{2} \|w\|^2$ we get ridge regression, when $f_i(w) = \log(1 + \exp(-b_i \langle w, x_i \rangle)) + \frac{\lambda}{2} \|w\|^2$ we get ℓ_2 regularized logistic regression, other examples such as SVMs can also be obtained by setting different individual functions $f_i(w)$. The ubiquitousness of such finite-sum optimization problems, and the massive scale of modern datasets motivate a lot of research effort on efficient optimization algorithms to solve (1). Let $w^* = \arg \min_w f(w)$ to be the optimum of (1), for any solution w we called it achieves ϵ -objective suboptimality if $f(w) - f(w^*) \leq \epsilon$. For any $\epsilon > 0$, the runtime cost of optimization algorithms to find ϵ -suboptimal minimizer typically depends on the target accuracy ϵ , as well as the conditions of the problem (1). Throughout the paper, we use the following notion of strong convexity and smoothness to discuss the iteration complexity of various optimization algorithms.

Definition 1. A function $f_i(w)$ is L -smooth with respect to w if $f_i(w)$ is differentiable and its gradient is L -Lipschitz continuous, i.e. we have

$$\|\nabla f_i(w) - \nabla f_i(w')\| \leq L \|w - w'\|, \quad \forall w, w' \in \mathbb{R}^d,$$

one consequence of L -smoothness is the following quadratic upper bound:

$$f(w) \leq f(w') + \langle \nabla f(w'), w - w' \rangle + \frac{L}{2} \|w - w'\|^2, \quad \forall w, w' \in \mathbb{R}^d.$$

A function $f_i(w)$ is λ -strongly convex with respect to w , if we have

$$f(w) \geq f(w') + \langle \nabla f(w'), w - w' \rangle + \frac{\lambda}{2} \|w - w'\|^2, \quad \forall w, w' \in \mathbb{R}^d.$$

Recent years witnessed a lot of advances in developing fast optimization algorithms for (1), we refer the readers to (Bottou et al., 2016) for a comprehensive survey of these developments. For large-scale problems in form of (1), randomized methods are particularly efficient because of its low per iteration cost. Below we briefly review two lines of research: i) randomized variance reduced first-order methods, ii) randomized methods leveraging second-order information. Through out this section we focus on problems where each function $f_i(w), \forall i \in [n]$ is λ -strongly convex and L -smooth.

Variance reduced first-order methods The key technique for developing fast stochastic first-order methods is variance reduction, which make sure the variance of the randomized updating direction decreases when the iterate gets closer to optimum. Representative methods of this category include SAG (Roux et al., 2012), SVRG (Johnson and Zhang, 2013), SDCA (Shalev-Shwartz and Zhang, 2013) and SAGA (Defazio et al., 2014), etc. These methods requires

$$\mathcal{O} \left(\left(n + \frac{L}{\lambda} \right) \log \left(\frac{1}{\epsilon} \right) \right) \quad (2)$$

calls to the first-order oracle of individual functions to find a solution that reaches ϵ -suboptimality. While in the large condition number regime (e.g. $\frac{L}{\lambda} > n$), using several acceleration techniques (e.g. Catalyst (Shalev-Shwartz and Zhang, 2016; Frostig et al., 2015; Lin et al., 2015a), APCG (Lin et al., 2015b), SPDC (Zhang and Xiao, 2015), Katyusha (Allen-Zhu, 2016), etc), one can further improve the iteration complexity to

$$\mathcal{O} \left(\left(n + \sqrt{n \cdot \frac{L}{\lambda}} \right) \log \left(\frac{1}{\epsilon} \right) \right). \quad (3)$$

Recently, (Woodworth and Srebro, 2016) provide first-order lower bound for (1) and showed the iteration complexity of (3), given first order and prox oracle of individual functions, is unimprovable for general (high dimensional) problems.

Leveraging second-order information Second-order information are often useful in improving the convergence of optimization algorithms. However for large-scale problems, obtaining and inverting the exact Hessian matrix is often computational expensive, which makes vanilla Newton methods not well suited in solving (1). There have been emerging studies in designing randomized algorithms which effectively utilize the *approximated* Hessian information. One lines of research is the sub-sampled Newton method which approximate the Hessian matrix based on a sub-sampled minibatch. (Byrd et al., 2011; Erdogdu and Montanari, 2015; Roosta-Khorasani and Mahoney, 2016a,b; Bollapragada et al., 2016) established local linear convergence rate for several variant of sub-sampled Newton methods, provided the sampling sizes is large enough. (Xu et al., 2016) considered non-uniform sampling in constructing the stochastic Hessian matrix, and showed that

weighted sampling according to individual smoothness or leverage score can help constructing better approximation of Hessian, thus improving the convergence. (Pilanci and Wainwright, 2017) discussed how to use sketching instead of sub-sampling to approximate the Hessian matrix. However, aforementioned methods are often combined with full gradient information, and require calling the second-order oracle as well as solving a related linear system at every iteration. Thus the computational complexity (as compared in Table 2 of (Xu et al., 2016)) are often worse than SVRG type algorithms.

Another line of research is to consider randomness in both first and second-order information to design lower per-iteration cost algorithms. In particular, (Byrd et al., 2016) considered combining minibatch SGD with Limited-memory BFGS (L-BFGS) (Liu and Nocedal, 1989) type update, and inspired by this, (Moritz et al., 2016; Gower et al., 2016; Wang et al., 2017b) proposed to combine variance reduced stochastic gradient with L-BFGS, and proved linear convergence for strongly convex and smooth objectives, (Moritz et al., 2016; Gower et al., 2016; Wang et al., 2017b) also demonstrated superior empirical performance of this type of methods. However, theoretically it is hard to guarantee the quality of approximated Hessian using L-BFGS update, thus the iteration complexity obtained by (Moritz et al., 2016; Gower et al., 2016; Wang et al., 2017b) can be pessimistic and being much worse than vanilla SVRG. Moreover, (Qu et al., 2015) proposed to incorporate curvature information in minibatch SDCA methods, and showed improved convergence over SDCA, but the method proposed in (Qu et al., 2015) involves solving a much more expensive subproblem than minibatch SDCA, thus the overall runtime benefit is still unclear. (Gonen et al., 2016) considered ridge regression problems specifically, and propose to use sketching to compute the rank- k approximation of the Hessian matrix, based on which SVRG is ran on a preconditioned space. (Yang et al., 2016) suggested to use preconditioning as a preprocessing step for general stochastic optimization problems. (Agarwal et al., 2016) proposed to approximate the inverse Hessian matrix directly by sampling from its Taylor expansion. In terms of lower bound, (Arjevani and Shamir, 2016) showed under some mild algorithmic assumptions, second-order oracle generally cannot improve the oracle complexity over first-order methods, for very high-dimensional problems.

Though lower bound have been established showing that the iteration complexity of accelerated SVRG methods cannot be improved in general, the establishment of these lower bounds are based on constructing very high-dimensional hard problems where d can be much larger than n . While in practice, second-order information are observed to be very helpful in improving the convergence speed, so it is still interesting to analyze theoretically how second-order information can be helpful in certain situations, and more importantly, how to design more efficient methods that make use of possibly noisy second-order information. We take effort in this direction and make the following contributions in this paper:

- We proposed a novel approach that combines the advantages of variance-reduced first-order methods and sub-sampled Newton methods, in a efficient way that does not require expensive Hessian matrix computation and inversion. The method can naturally be extended to also solve composite optimization problems with non-smooth regularization.
- We theoretically show under certain conditions the proposed approach can improve state-of-the-art iteration complexity, and empirically demonstrated it can substantially boost the convergence of existing methods.

Organization The rest of the paper is organized as follows: we introduce the proposed method in Section 2, and present the main theoretical results in Section 3, in Section 4 we present the convergence analysis inexact minibatch accelerated SVRG update as a key step in proving the

main results, this might be of independent interest. We provide some empirical comparisons over existing approaches in Section 5, and conclude in Section 6. Some detailed proofs are deferred to Appendix.

Notations We use $[n]$ to denote the set $1, \dots, n$. For a vector $w \in \mathbb{R}^d$, we use $\|w\|$ to denote its ℓ_2 norm and $\|w\|_1$ to denote its ℓ_1 norm. For a matrix X , we use $\|X\|_2$ to denote its spectral norm and $\|X\|_F$ to denote its Frobenius norm, and $\lambda_{\min}(X)$ to denote its minimum singular value, and $\text{tr}(X)$ to denote the trace of X . We use I to denote an identity matrix. For two symmetric matrices A and B , we denote $A \succeq B$ if $A - B$ is positive semi-definite. For two sequences of numbers $\{a_n\}$ and $\{b_n\}$, we say $a_n = \mathcal{O}(b_n)$ if $a_n \leq Cb_n$ for n large enough, with some positive constant C , we use the notation $\tilde{\mathcal{O}}(\cdot)$ to hide poly-log factors. We also use $a_n \lesssim b_n$ to denote $a_n = \mathcal{O}(b_n)$, $a_n \gtrsim b_n$ to denote $b_n = \mathcal{O}(a_n)$, $a_n \asymp b_n$ to denote $a_n = \mathcal{O}(b_n)$ and $b_n = \mathcal{O}(a_n)$.

2 Minibatch Stochastic Variance Reduced Proximal Iterations

In this section we present the proposed approach for minimizing (1), which naturally integrate both noisy first-order and higher-order information in an efficient way. We first propose and discuss the main building block, which we called the minibatch stochastic variance reduced proximal iterations, then present the complete algorithm. Suppose at iteration t , given the previous iterate w_{t-1} , we consider the following update rule:

$$w_t = \arg \min_w \frac{1}{b} \sum_{i \in \bar{B}} f_i(w) - \left\langle \frac{1}{b} \sum_{i \in \bar{B}} \nabla f_i(w_{t-1}), w \right\rangle + \left\langle \frac{\eta}{b} \sum_{i \in B_t} \nabla f_i(w_{t-1}) + \eta \nabla f(\tilde{w}) - \frac{\eta}{b} \sum_{i \in B_t} \nabla f_i(\tilde{w}), w \right\rangle + \frac{\tilde{\lambda}}{2} \|w - w_{t-1}\|^2, \quad (4)$$

where \bar{B}, B_t are some randomly sampled minibatch from $1, \dots, n$, both with minibatch size b , and $\eta, \tilde{\lambda}$ are stepsize parameters, \tilde{w} is a “reference” predictor, used to reduce the variance. Depending on the choice of these parameters, we observe that the update rule of (4) can be viewed as a generalization of several update rules proposed recently:

- When $b = 1$, $\tilde{\lambda} \rightarrow \infty$ and $\eta = \frac{\tilde{\lambda}}{L}$, the term $\frac{1}{b} \sum_{i \in \bar{B}} f_i(w) - \langle \frac{1}{b} \sum_{i \in \bar{B}} \nabla f_i(w_{t-1}), w \rangle$ is negligible, thus (4) reduced to standard SVRG update (Johnson and Zhang, 2013):

$$w_t \leftarrow w_{t-1} - \frac{1}{L} \left(\nabla f_i(w_{t-1}) + \frac{1}{n} \sum_{i=1}^n \nabla f_i(\tilde{w}) - \nabla f_i(\tilde{w}) \right).$$

- When $b > 1$, $\tilde{\lambda}, \eta \rightarrow \infty$, then the term $\frac{1}{b} \sum_{i \in \bar{B}} f_i(w) - \langle \frac{1}{b} \sum_{i \in \bar{B}} \nabla f_i(w_{t-1}), w \rangle$ in (4) is negligible as well, (4) reduces to the following update

$$w_t \leftarrow w_{t-1} - \frac{\eta}{\tilde{\lambda}} \left(\frac{1}{b} \sum_{i \in B_t} \nabla f_i(w_{t-1}) + \frac{1}{n} \sum_{i=1}^n \nabla f_i(\tilde{w}) - \frac{1}{b} \sum_{i \in B_t} \nabla f_i(\tilde{w}) \right),$$

which recovers the update rule of minibatch semi-stochastic gradient methods (a.k.a minibatch SVRG) (Konečný et al., 2016).

- When $\bar{B} = B_t$, $b = 1$, $\eta = 1$, (4) reduced to stochastic variance reduced proximal iterations. More specifically, (4) will be reduced to

$$w_t = \arg \min_w f_i(w) + \left\langle \frac{1}{n} \sum_{i=1}^n \nabla f_i(\tilde{w}) - \nabla f_i(\tilde{w}), w \right\rangle + \frac{\tilde{\lambda}}{2} \|w - w_{t-1}\|^2, \quad (5)$$

by checking the first order optimality condition it is easy to verify that (5) is performing the following update:

$$w_t \leftarrow w_{t-1} - \frac{1}{\tilde{\lambda}} \left(\nabla f_i(w_t) + \frac{1}{n} \sum_{i=1}^n \nabla f_i(\tilde{w}) - \nabla f_i(\tilde{w}) \right),$$

compared with standard SVRG update with stepsize $\frac{1}{\tilde{\lambda}}$, we see the only difference is the gradient evaluation on f_i is based on “future” iterate w_t rather than the current iterate w_{t-1} in SVRG. Stochastic proximal iterations based on SAGA method has been analyzed in (Defazio, 2016) recently, and shown to achieve accelerated rate without explicit momentum step.

- When $B_t = \bar{B}$, $b > 1$, $\eta = 1$, and $\tilde{w} = w_{t-1}$, (4) will be reduced to

$$w_t = \arg \min_w \frac{1}{b} \sum_{i \in \bar{B}} f_i(w) + \left\langle \frac{1}{n} \sum_{i=1}^n \nabla f_i(w_{t-1}) - \frac{1}{b} \sum_{i \in \bar{B}} \nabla f_i(w_{t-1}), w \right\rangle + \frac{\tilde{\lambda}}{2} \|w - w_{t-1}\|^2, \quad (6)$$

(6) covers the update rule of DANE algorithm (Shamir et al., 2014), which is a communication-efficient distributed optimization algorithm. DANE uses the data on local machine to form the minibatch \bar{B} and every round machines communicate the gradient vector based on their local data. As shown in (Shamir et al., 2014), DANE is provably communication more efficient than the first-order methods in certain scenarios.

- When $\bar{B} = B_t$, and we ignore the linear term $-\langle \frac{1}{b} \sum_{i \in \bar{B}} \nabla f_i(w_{t-1}), w \rangle + \langle \frac{\eta}{b} \sum_{i \in B_t} \nabla f_i(w_{t-1}) + \eta \nabla f(\tilde{w}) - \frac{\eta}{b} \sum_{i \in B_t} \nabla f_i(\tilde{w}), w \rangle$, then (4) reduces to the minibatch proximal iterations (Li et al., 2014; Wang et al., 2017a):

$$w_t = \arg \min_w \frac{1}{b} \sum_{i \in B_t} f_i(w) + \frac{\tilde{\lambda}}{2} \|w - w_{t-1}\|^2.$$

Such an update allows larger minibatch size than standard minibatch SGD, but without the linear correction term as we considered in (4), only sublinear convergence can be established for finite-sum problems.

The main difference with these known methods is (4) considered the information brought by the term $\frac{1}{b} \sum_{i \in \bar{B}} f_i(w) - \langle \frac{1}{b} \sum_{i \in \bar{B}} \nabla f_i(w_{t-1}), w \rangle$, which incorporate noisy higher-order information of the functions. In particular, second-order methods (e.g. Newton methods) often use the following second-order Taylor approximation of the function:

$$\begin{aligned} \frac{1}{b} \sum_{i \in \bar{B}} f_i(w) &\approx \frac{1}{b} \sum_{i \in \bar{B}} f_i(w_{t-1}) + \left\langle \frac{1}{b} \sum_{i \in \bar{B}} \nabla f_i(w_{t-1}), w - w_{t-1} \right\rangle \\ &\quad + \frac{1}{2} (w - w_{t-1})^\top \left(\frac{1}{b} \sum_{i \in \bar{B}} \nabla^2 f_i(w_{t-1}) \right) (w - w_{t-1}), \end{aligned} \quad (8)$$

Algorithm 1 MB-SVRP: Minibatch Stochastic Variance Reduced Proximal Iterations.

Parameters $\eta, b, \tilde{\lambda}, \nu, \varepsilon$.

Initialize $\tilde{w}_0 = 0$.

Sampling Sampling b items from $[n]$ to form a minibatch \bar{B} .

for $s = 1, 2, \dots$ **do**

Calculate $\tilde{v} = \frac{1}{n} \sum_{i=1}^n \nabla f_i(\tilde{w}_{s-1})$.

Initialize $y_0 = w_0 = \tilde{w}_s$.

for $t = 1, 2, \dots, m$ **do**

Sampling b items from $[n]$ to form a minibatch B_t .

Approximately solve the minimization problem to get w_t (using SVRG or other variants):

$$w_t \approx \arg \min_w \tilde{f}_t(w) := \frac{1}{b} \sum_{i \in \bar{B}} f_i(w) - \left\langle \frac{1}{b} \sum_{i \in \bar{B}} \nabla f_i(y_{t-1}), w \right\rangle + \left\langle \frac{\eta}{b} \sum_{i \in B_t} \nabla f_i(y_{t-1}) + \eta \tilde{v} - \frac{\eta}{b} \sum_{i \in B_t} \nabla f_i(\tilde{w}_{s-1}), w \right\rangle + \frac{\tilde{\lambda}}{2} \|w - y_{t-1}\|^2, \quad (7)$$

such that

$$w_t - \min_w \tilde{f}_t(w) \leq \varepsilon.$$

Update:

$$y_t = w_t + \nu(w_t - w_{t-1}).$$

end for

Update $\tilde{w}_s = w_m$.

end for

Return \tilde{w}_s

if we replace the second-order approximation (8) into (4), we get the update rule as

$$\begin{aligned} w_t &= \arg \min_w \frac{1}{2} (w - w_{t-1})^\top \left(\frac{1}{b} \sum_{i \in \bar{B}} \nabla^2 f_i(w_{t-1}) \right) (w - w_{t-1}) + \frac{\tilde{\lambda}}{2} \|w - w_{t-1}\|^2 \\ &\quad + \left\langle \frac{\eta}{b} \sum_{i \in B_t} \nabla f_i(w_{t-1}) + \eta \nabla f(\tilde{w}) - \frac{\eta}{b} \sum_{i \in B_t} \nabla f_i(\tilde{w}), w \right\rangle \\ &= w_{t-1} - \eta \left(\frac{1}{b} \sum_{i \in \bar{B}} \nabla^2 f_i(w_{t-1}) + \tilde{\lambda} I \right)^{-1} \left(\frac{1}{b} \sum_{i \in B_t} \nabla f_i(w_{t-1}) + \nabla f(\tilde{w}) - \frac{1}{b} \sum_{i \in B_t} \nabla f_i(\tilde{w}) \right), \quad (9) \end{aligned}$$

which can be treated as a variant of sub-sampled Newton method, combined with the minibatch stochastic gradient with variance reduction. Moreover, when $f_i(w)$ is a quadratic function of w , the approximation (8) is exact, thus the update rule (4) can be treated as a preconditioned minibatch SVRG update rule, with $\left(\frac{1}{b} \sum_{i \in \bar{B}} \nabla^2 f_i(w_{t-1}) + \tilde{\lambda} I \right)^{-1}$ as a precondition matrix. This is formalized in the proposition below.

Proposition 2. *When $f_i(w), \forall i$ is a quadratic function of w , then the update rule of (4) is equiv-*

alent to the following preconditioned minibatch svrg update rule:

$$w_t \leftarrow w_{t-1} - \eta \left(\tilde{H} + \tilde{\lambda} I \right)^{-1} \left(\frac{1}{b} \sum_{i \in B_t} \nabla f_i(w_{t-1}) + \nabla f(\tilde{w}) - \frac{1}{b} \sum_{i \in B_t} \nabla f_i(\tilde{w}) \right),$$

where

$$\tilde{H} = \frac{1}{b} \sum_{i \in \bar{B}} \nabla^2 f_i(w)$$

is the sub-sampled Hessian matrix.

For general, non-quadratic functions $f_i(w)$ where $\nabla^2 f_i(w)$ is changing with w , hopefully that using the exact function as in (4), which takes higher-order information into consideration, thus being not worse, or even better than the quadratic approximation considered in (9). Moreover, using exact local functions instead of second-order approximation also brings important computational advantage since we can solve (4) without requiring the second-order oracle and Hessian matrix inverting operations, which are often computational expensive in large-scale problems.

Using (4) as building blocks, we propose the MB-SVRP (minibatch stochastic variance reduced proximal iterations) method, which is detailed in Algorithm 1¹. At the beginning of the algorithm, we form a minibatch \bar{B} by sampling from $1, \dots, n$ and fix it for the whole optimization process², then following the SVRG method (Johnson and Zhang, 2013), the algorithm is divided to multiple stages, indexed by s , at each stage, we iteratively solve a minimization problem of the form (7) based on the randomly sampled minibatch B_t , the difference with (4) is that we consider a momentum scheme by maintaining two sequences $\{w_t, y_t\}$, which is inspired by Nesterov's acceleration technique (Nesterov, 2004) and its recent SVRG variant (Nitanda, 2014). The main theoretical advantage compared with ones without momentum (Konečný et al., 2016) is by such an acceleration step we are able to use a much larger minibatch size without slowing down the convergence, while standard minibatch SVRG can only tolerate constant minibatch size (Konečný et al., 2016). The advantage of allowing larger minibatch size using acceleration in SVRG type algorithms is in analogy to the situation in stochastic gradient descent (without variance reduction) type algorithms (Dekel et al., 2012; Cotter et al., 2011; Lan, 2012). Finally, since its often expensive find the exact minimizer of (7), here we consider an approximate minimizer with objective suboptimality ε . When we choose the appropriate $\tilde{\lambda}$ for $\tilde{f}_t(w)$ to obtain enough strong convexity, (7) can be solved efficiently using recent advances of finite-sum optimization (e.g. SVRG) even for relatively small ε . Allowing error in gradient oracle have been analyzed in several batch first-order methods (Schmidt et al., 2011; Devolder et al., 2014), but being largely unexplored in stochastic gradient methods, except for recent work of (Wang et al., 2017a) which analyzed inexact minibatch prox update. In Section 4, we establish convergence rate of *inexact, minibatched, accelerated* SVRG method, which might be of independent interest.

2.1 Extension to composite minimization

For many methods try to incorporate second-order information such as sub-sampled Newton and L-BFGS type algorithms, it is not clear how to extend them to solve non-smooth composite prob-

¹In step (7), as discussed above it is also possible to just use second-order approximation instead of the exact function $\frac{1}{b} \sum_{i \in \bar{B}} f_i(w)$, as (8) did, then the method can be viewed as a mix of sub-sampled Newton method and SVRG.

²We can also consider the varying \bar{B} option by simply setting $\bar{B} = B_t$, which makes the algorithm simpler to implement and in practice, we observed no significant difference with the pre-fixed \bar{B} , here we consider fixed \bar{B} mainly for the sake of simplicity in theoretical analysis.

lems. In contrast, the proposed approach can be easily extended to solve non-smooth composite minimization problems as well. Consider minimization of:

$$F(w) := \frac{1}{n} \sum_{i=1}^n f_i(w) + g(w), \quad (10)$$

where the component function $f_i(w)$ is smooth and strongly convex, but we also considered a non-smooth regularization. For example when $g(w) = \mu \|w\|_1$ and $f_i(w) = \frac{1}{2}(b_i - \langle w, x_i \rangle)^2$ we get the Lasso objective (Tibshirani, 1996), when $g(w) = \mu \|w\|_1 + \frac{\lambda}{2} \|w\|^2$ and $f_i(w) = \log(1 + \exp(-b_i \langle w, x_i \rangle))$ we get the elastic net regularized logistic regression (Zou and Hastie, 2005).

We modify the Algorithm 1 to solve (10). The idea is rather straightforward: at each inner iteration, rather than (7), instead we simply solve the following minibatch composite minimization problem:

$$\begin{aligned} w_t \approx \arg \min_w \tilde{F}_t(w) := & \frac{1}{\eta b} \sum_{i \in \bar{B}} f_i(w) - \left\langle \frac{1}{\eta b} \sum_{i \in \bar{B}} \nabla f_i(y_{t-1}), w \right\rangle \\ & + \left\langle \frac{1}{b} \sum_{i \in B_t} \nabla f_i(y_{t-1}) + \tilde{v} - \frac{1}{b} \sum_{i \in B_t} \nabla f_i(\tilde{w}_{s-1}), w \right\rangle + \frac{\tilde{\lambda}}{2\eta} \|w - y_{t-1}\|^2 + g(w), \end{aligned} \quad (11)$$

here we slightly re-scale the approximated loss term to make sure the relative weight between approximated loss and regularization is correct. Since the objective (11) is a standard finite-sum composite minimization problem we could apply prox-SVRG or prox-SAGA to solve (11) efficiently when the term $\frac{\tilde{\lambda}}{2\eta} \|w - y_{t-1}\|^2$ brings sufficient strong convexity.

3 Convergence analysis for quadratic objectives

In this section we present theoretical results for the proposed approach. We focused on analysis for quadratic objectives for which we are able to show improved convergence under reasonable assumptions. As discussed in Section 2, several known algorithms (such as minibatch SVRG (Konečný et al., 2016)) can be covered by the proposed MB-SVRP method, thus for general convex objectives, the convergence analysis of these algorithms can also be applied for MB-SVRP as well by choosing the corresponding parameters (\bar{B} , B_t , η , b , and $\tilde{\lambda}$). Moreover, when we use the second-order approximation option as discussed in (9), the algorithm can be viewed as a combination of minibatch SVRG and sub-sampled Newton, theoretical analysis of sub-sampled Newton methods can be extended to analyze this option as well.

For this section we focused on the problems when $f_i(w)$ is quadratic (a.k.a ridge regression):

$$\min_w \frac{1}{n} \sum_{i=1}^n f_i(w) := \frac{1}{n} \sum_{i=1}^n \frac{(y_i - x_i^\top w)^2}{2} + \frac{\lambda}{2} \|w\|^2 = \frac{1}{n} \sum_{i=1}^n \left(\frac{(y_i - x_i^\top w)^2}{2} + \frac{\lambda}{2} \|w\|^2 \right), \quad (12)$$

where one can treat the component functions $f_i(w)$ as $f_i(w) = \frac{(y_i - x_i^\top w)^2}{2} + \frac{\lambda}{2} \|w\|^2$, in such case, it is clear that the individual smoothness parameter of the loss can be upper bounded by

$$L = \max_{i \in [n]} \|x_i\|^2,$$

while the strong convexity parameter can be lower bounded by λ .

Let $H_\lambda = \frac{1}{n} \sum_{i=1}^n x_i x_i^\top + \lambda I$ to be the Hessian matrix, and suppose we have an approximated Hessian $\tilde{H}_{\tilde{\lambda}}$, the following proposition states that performing the preconditioned minibatch SVRG update is equivalent to use standard minibatch SVRG in a linear transformed (preconditioned) space:

Proposition 3. *Considering the following minibatch SVRG type update when solving $\min_w f(w)$:*

$$w_t \leftarrow w_t - \eta \tilde{H}_{\tilde{\lambda}}^{-1} \left(\frac{1}{b} \sum_{i \in B_t} \nabla f_i(w_{t-1}) + \frac{1}{n} \sum_{i=1}^n \nabla f_i(\tilde{w}) - \frac{1}{b} \sum_{i \in B_t} \nabla f_i(\tilde{w}) \right), \quad (13)$$

it is equivalent to solving a minimization problem with respect to z : $\min_z f(\tilde{H}_{\tilde{\lambda}}^{-1/2} z)$ via the exchange of variables: $z = \tilde{H}_{\tilde{\lambda}}^{1/2} w$. Moreover, the update rule (13) is equivalent to the performing the following update on z :

$$z_t \leftarrow z_t - \eta \left(\frac{1}{b} \sum_{i \in B_t} \nabla_z f_i(\tilde{H}_{\tilde{\lambda}}^{-1/2} z_{t-1}) + \frac{1}{n} \sum_{i=1}^n \nabla_z f_i(\tilde{H}_{\tilde{\lambda}}^{-1/2} \tilde{z}) - \frac{1}{b} \sum_{i \in B_t} \nabla_z f_i(\tilde{H}_{\tilde{\lambda}}^{-1/2} \tilde{z}) \right), \quad (14)$$

With above proposition, we see that for quadratic objectives, the proposed update (4) is implicitly performing minibatch SVRG update on the following transformed problem:

$$\min_z f(\tilde{H}_{\tilde{\lambda}}^{-1/2} z) := \frac{1}{2} z^\top \tilde{H}_{\tilde{\lambda}}^{-1/2} \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^\top + \lambda I \right) \tilde{H}_{\tilde{\lambda}}^{-1/2} z - \left\langle \frac{X^\top y}{n}, \tilde{H}_{\tilde{\lambda}}^{-1/2} z \right\rangle + \frac{\|y\|^2}{2n}, \quad (15)$$

as a straightforward extension, it is also easy to see when combining the acceleration technique (as Algorithm 1 did), is equivalent to perform *accelerated* minibatch SVRG update, as analyzed in (Nitanda, 2014). Hopefully when $\tilde{H}_{\tilde{\lambda}}$ is close to H_λ , the condition number of the new problem (15) becomes smaller than that of the original problem (12), as a consequence, when performing the update (13) is cheap we would expect an improved convergence and runtime guarantees for solving (12). Before analyzing the condition number, we state the following lemma which connects the inexactness in w space when performing the update (4) (and (13)) to the inexactness in z space when performing the update (14).

Lemma 4. *Let $\tilde{f}_t(w)$ be the function to be minimized in (7), and $\bar{w}_t = \arg \min_w \tilde{f}_t(w)$ to be its exact minimizer. For any w_t satisfies*

$$\tilde{f}_t(w_t) - \tilde{f}_t(\bar{w}_t) \leq \varepsilon,$$

we must have the following bound on $z_t - \bar{z}_t$:

$$\|z_t - \bar{z}_t\| \leq \sqrt{\frac{2\varepsilon(L + \tilde{\lambda})}{\lambda + \tilde{\lambda}}},$$

where $z_t = \tilde{H}_{\tilde{\lambda}}^{1/2} w_t$ and $\bar{z}_t = \tilde{H}_{\tilde{\lambda}}^{1/2} \bar{w}_t$.

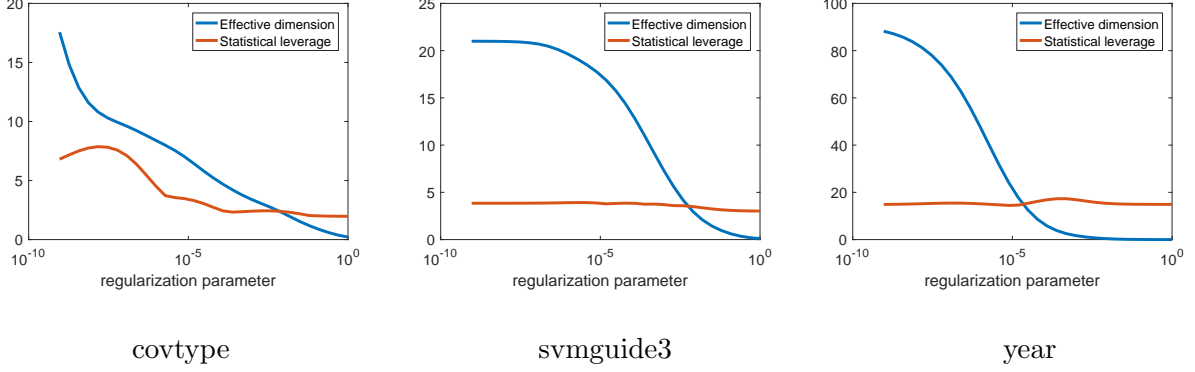


Figure 1: Effective dimension and maximum statistical leverage with different regularization parameters $\tilde{\lambda}$ on some empirical datasets.

3.1 Improved condition number

In this section we analyze the condition number in the “preconditioned” space (15), where \bar{B} is sampled uniformly from $1, \dots, n$. For the analysis, we introduce two notions which describes the global and local properties of data. The following definitions of effective dimension and bounded statistical leverage can be found (Hsu et al., 2014) which are used to analyzing the generalization performance of ridge regression.

Definition 5. (Effective dimension) Let the $\lambda_1, \dots, \lambda_d$ be the top- d eigenvalues of $\frac{1}{n} \sum_{i=1}^n x_i x_i^\top$, define the effective dimension $d_{\tilde{\lambda}}$ (for some $\tilde{\lambda} \geq 0$) of $\frac{1}{n} \sum_{i=1}^n x_i x_i^\top$ being

$$d_{\tilde{\lambda}} = \sum_{j=1}^d \frac{\lambda_j}{\lambda_j + \tilde{\lambda}}.$$

We see $d_{\tilde{\lambda}}$ is a decreasing function of $\tilde{\lambda}$, more specifically when $\tilde{\lambda} = 0$, then $d_0 = d$. When the spectrum of $\frac{1}{n} \sum_{i=1}^n x_i x_i^\top$ is decaying very fast, the effective dimension $d_{\tilde{\lambda}}$ can be significantly smaller than d for moderate $\tilde{\lambda}$. The following notion of statistical leverage have been used in regression analysis (Chatterjee and Hadi, 2009; Hsu et al., 2014) and matrix approximation (Mahoney et al., 2011).

Definition 6. (Statistical leverage at $\tilde{\lambda}$) Let $H_{\tilde{\lambda}} = \frac{1}{n} \sum_{i=1}^n x_i^\top x_i + \tilde{\lambda}I$, we say the statistical leverage of data matrix X is bounded by $\rho_{\tilde{\lambda}}$ at $\tilde{\lambda}$ if

$$\frac{\|H_{\tilde{\lambda}}^{-1/2} x_i\|}{\sqrt{(1/n) \sum_{j=1}^n \|H_{\tilde{\lambda}}^{-1/2} x_j\|^2}} \leq \rho_{\tilde{\lambda}}, \quad (16)$$

hold for every $i \in [n]$.

Above definition is slightly different from one used in (Hsu et al., 2014) in the sense that the empirical Hessian matrix $H_{\tilde{\lambda}}$ in (16) is replaced by the population Hessian matrix $\mathbb{E}[x_i^\top x_i] + \tilde{\lambda}I$ in (Hsu et al., 2014), when the sample size n is large, the differences in these two definitions are

minor. As argued in (Hsu et al., 2014), when x_i is drawn from subgaussian distributions, then $(\mathbb{E}[x_i^\top x_i])^{-1/2} x_i$ is isotropic, in which case the statistical leverage only grows logarithmically with the dimension. Our theoretical analysis relies on above two defined quantities are not too large, Figure 1 showed the effective dimension and statistical leverage on several real world datasets with varying regularization parameters.

Suppose we construct $\tilde{H}_{\tilde{\lambda}}$

$$\tilde{H}_{\tilde{\lambda}} = \frac{1}{b} \sum_{i \in \bar{B}} x_i x_i^\top + \tilde{\lambda} I,$$

where \bar{B} is a batch of size b sampled uniformly from $[n]$. For stochastic gradient based algorithms in minimizing the objective (23), the relevant two key quantities are strong convexity and smoothness. The strong convexity for problem (23) is $\lambda_{\min}(\tilde{H}_{\tilde{\lambda}}^{-1/2} (\frac{1}{n} \sum_{i=1}^n x_i x_i^\top + \lambda I) \tilde{H}_{\tilde{\lambda}}^{-1/2})$, which is a global property of the objective; while for smoothness, since we are considering using variance reduced stochastic gradient methods in the preconditioned space, we must consider the smoothness parameter for individual function $f_i(\tilde{H}_{\tilde{\lambda}}^{-1/2} z)$, which in our context, is $\max_i \{x_i^\top \tilde{H}_{\tilde{\lambda}}^{-1} x_i\}$. Both of these two key quantities are closed related to how close is the constructed Hessian approximation $\tilde{H}_{\tilde{\lambda}}$, to the true Hessian $H_{\tilde{\lambda}}$ in spectral norm. The following lemma bound $\|H_{\tilde{\lambda}}^{-1} (\frac{1}{b} \sum_{i \in \bar{B}} x_i x_i^\top - \frac{1}{n} \sum_{i=1}^n x_i x_i^\top)\|_2$ using matrix concentration.

Lemma 7. *If \bar{B} is formed by uniform sampling with replacement from $[n]$, then we have the following concentration bound, with probability at least $1 - \delta$,*

$$\left\| H_{\tilde{\lambda}}^{-1} \left(\frac{1}{b} \sum_{i \in \bar{B}} x_i x_i^\top - \frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right) \right\|_2 \leq 2 \log \left(\frac{d}{\delta} \right) \cdot \sqrt{\frac{\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}}}{b}}.$$

Strong Convexity Based on above lemma, we have the following lower bound of the strong convexity for (15), specified in the following lemma.

Lemma 8. *If we set $\tilde{\lambda}$ such that $\tilde{\lambda} \geq \lambda$, then with probability at least $1 - \delta$ over the random choice of \bar{B} to form $\tilde{H}_{\tilde{\lambda}}$, we have*

$$\lambda_{\min}(\tilde{H}_{\tilde{\lambda}}^{-1/2} H_{\lambda} \tilde{H}_{\tilde{\lambda}}^{-1/2}) \geq \frac{\lambda}{\tilde{\lambda}} \frac{1}{1 + 2 \log(d/\delta) \sqrt{(\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}})/b}}.$$

Smoothness Next we explore the smoothness parameter of (15), of which the goal is to upper bound

$$\max_{i \in [n]} x_i^\top \tilde{H}_{\tilde{\lambda}}^{-1} x_i.$$

The most straightforward way to upper bound $\max_{i \in [n]} x_i^\top \tilde{H}_{\tilde{\lambda}}^{-1} x_i$ is

$$\max_{i \in [n]} x_i^\top \tilde{H}_{\tilde{\lambda}}^{-1} x_i \leq \max_{i \in [n]} \|x_i\|^2 \lambda_{\min}(\tilde{H}_{\tilde{\lambda}}^{-1}) = \max_{i \in [n]} \frac{\|x_i\|^2}{\tilde{\lambda}} = \frac{L}{\tilde{\lambda}},$$

by that way, we get the condition number after preconditioning is

$$\mathcal{O}\left(\frac{4\tilde{\lambda}}{\lambda} \cdot \frac{L}{\tilde{\lambda}}\right) = \frac{4L}{\lambda},$$

which didn't show advantage after preconditioning. In the lemma below, we provide an improved analysis which based on the notion of effective dimension (Definition 5) and statistical leverage (Definition 6), which is specified in the lemma blow.

Lemma 9. *If we choose $\tilde{\lambda} \geq \lambda$ and $b \geq 16\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}} \log^2(d/\delta)$, then with probability at least $1 - \delta$ over the random choice of \bar{B} to form $\tilde{H}_{\tilde{\lambda}}$, we have*

$$\max_{i \in [n]} x_i^\top \tilde{H}_{\tilde{\lambda}}^{-1} x_i^\top \leq 2\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}}$$

Combining Lemma 8 and Lemma 9 we get the following corollary about the condition number for (15), if we choose $\tilde{\lambda}$ and b satisfies:

$$\tilde{\lambda} \geq \max\left\{\lambda, \frac{L}{b}\right\}, \quad b \geq 16\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}} \log^2\left(\frac{d}{\delta}\right),$$

then with probability at least $1 - \delta$, the condition number for stochastic gradient algorithms after preconditioning scales as $\frac{3\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}} \tilde{\lambda}}{\lambda}$, more specifically when $\tilde{\lambda} = \max\left\{\lambda, \frac{L}{b}\right\}$, then the condition number of (15) can be upper bounded by:

$$\max\left\{3\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}}, \frac{3L\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}}}{\lambda b}\right\},$$

which improves the original condition number L/λ by a factor of at least $\tilde{\mathcal{O}}(\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}}/b)$.

3.2 Improved runtime guarantee

Based on above analysis, we have the following main results stating the iteration complexity of MP-SVRG algorithms applied on ridge regression problems (12).

Theorem 10. *When applying (1) on ridge regression problems (12). If we sample uniformly to from $1, \dots, n$ form \bar{B} , set the parameter $\tilde{\lambda}$ as*

$$\tilde{\lambda} = \max\left\{\lambda, \frac{L}{b}\right\},$$

and choose the minibatch size as

$$b \asymp \min\left\{n, \rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}} \left(\frac{L}{\lambda}\right)^{1/3}\right\},$$

and if we set the stepsize parameter η as

$$\eta \asymp \min\left\{\frac{b^3 \lambda}{(\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}})^2 L}, \frac{1}{\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}}}\right\}.$$

Then for MP-SVRP algorithm to find \tilde{w}_s reach expected ϵ -objective suboptimality in solving (12),

$$\mathbb{E}f(\tilde{w}_s) - \min_w f(w) \leq \epsilon,$$

if use SVRG to solve each subproblem in (7) such that the suboptimality ε satisfies

$$\varepsilon \leq \frac{1}{10^5} \cdot \left(\frac{\lambda}{L}\right)^7 \epsilon,$$

then the total number of gradient evaluations used in the whole MP-SVRP algorithm can be upper bounded by

$$\mathcal{O}\left(\tilde{\kappa} \cdot \log\left(\frac{L}{\lambda}\right) \cdot \log^2\left(\frac{1}{\epsilon}\right) + n \cdot \log\left(\frac{1}{\epsilon}\right)\right),$$

where

$$\tilde{\kappa} = \max\left\{\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}} \left(\frac{L}{\lambda}\right)^{2/3}, \frac{\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}} L}{\lambda n}\right\}.$$

Remark 1. Use the fact that $L_s \leq L$, we get the simplified iteration complexity of Theorem 10 as

$$\tilde{\mathcal{O}}\left(n + \max\left\{\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}} \left(\frac{L}{\lambda}\right)^{2/3}, \frac{\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}} L}{\lambda n}\right\}\right),$$

where $\tilde{\mathcal{O}}(\cdot)$ hide the minor poly logarithmic factors. When $\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}}$ is small such that can be treated as constant, then the iteration complexity of MB-SVRP improves over standard SVRG by a factor of $\min\left\{n, \left(\frac{L}{\lambda}\right)^{1/3}\right\}$ when the condition number $\frac{L}{\lambda}$ is much larger than n . Even compared with accelerated methods (such as SVRG equipped with catalyst acceleration (Shalev-Shwartz and Zhang, 2016; Frostig et al., 2015; Lin et al., 2015a)), it can sometimes better (when $n < L/\lambda < n^2$, see Table 1) for details.

Remark 2. As shown above MB-SVRP method is able to improve the iteration complexity over SVRG when $\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}}$ is small, thus it is also possible to use MB-SVRP itself to solve the minibatch subproblem (7). Such a nested approach allows us to choose a even smaller $\tilde{\lambda}$, thus can further reduce the dependency on condition number in the iteration complexity, but at the cost of increasing the dependency on $\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}}$ and more complicated implementation, we leave such investigation to future research.

3.3 With catalyst acceleration

Theorem 10 stated that if the condition number is not too large: $\frac{L}{\lambda} \leq n^{5/4}$, then MB-SVRP only requires logarithmic passes over data to find a solution with high accuracy. When the condition number is large, MP-SVRP can still be slow. By using the accelerated proximal point framework proposed in (Shalev-Shwartz and Zhang, 2016; Frostig et al., 2015; Lin et al., 2015a), it is possible to obtain a accelerated convergence rate which has milder dependence on condition number, the algorithm is outlined in Algorithm 2, which iteratively, approximately call the original algorithm MB-SVRP to solve an augmented proximal point problem. The main iteration complexity is stated in the theorem below.

Algorithm 2 Acc-MB-SVRP: Accelerated MB-SVRP method.

Initialize $w_0 = z_0 = 0$.

for $r = 1, 2, \dots$ **do**

Call **MP-SVRP** algorithm 1 to approximately solve

$$w_r \approx \arg \min_w f(w) + \frac{\gamma}{2} \|w - z_{r-1}\|. \quad (17)$$

Update

$$z_r = w_r + \nu_r(w_r - w_{r-1}).$$

end for

Return w_r .

Table 1: Comparison of iteration complexity of various finite-sum quadratic optimization algorithms when effective dimension and statistical leverage are bounded, where we compare the different relative scale of condition number $\kappa = L/\lambda$ and sample size n , ignoring logarithmic factors.

	$\kappa \leq n$	$\kappa = n^{4/3}$	$\kappa = n^{3/2}$	$\kappa = n^2$	$\kappa = n^3$
SVRG	n	$n^{4/3}$	$n^{3/2}$	n^2	n^3
Acc-SVRG	n	$n^{7/6}$	$n^{5/4}$	$n^{3/2}$	n^2
MB-SVRP	n	$\rho_\lambda^2 d_\lambda \cdot n$	$\rho_\lambda^2 d_\lambda \cdot n$	$\rho_\lambda^2 d_\lambda \cdot n^{3/2}$	$\rho_\lambda^2 d_\lambda \cdot n^{5/2}$
Acc-MB-SVRP	n	$(\rho_\lambda^2 d_\lambda)^{1/2} \cdot n$	$(\rho_\lambda^2 d_\lambda)^{1/2} \cdot n$	$(\rho_\lambda^2 d_\lambda)^{1/2} \cdot n^{5/4}$	$(\rho_\lambda^2 d_\lambda)^{1/2} \cdot n^{7/4}$

Theorem 11. For ill-condition problems where $L/\lambda \geq n^{3/2}$, if we set the parameter $\gamma \asymp \frac{\rho_\lambda^2 d_\lambda L}{n^{3/2}} - \lambda$, Algorithm 2 has iteration complexity of

$$\tilde{O} \left((\rho_\lambda^2 d_\lambda)^{1/2} \cdot n^{1/4} \left(\frac{L}{\lambda} \right)^{1/2} \right).$$

Remark 3. From theorem 11 we see the iteration complexity of accelerated MP-SVRP has only grows at a square root rate with the condition number, which is much better than the non-accelerated algorithms, and being similar to accelerated SVRG algorithms. Moreover, accelerated MP-SVRP improves the iteration complexity of accelerated SVRG by a factor of $n^{1/4}$ when $\sqrt{\rho_\lambda^2 d_\lambda}$ is small.

4 Convergence analysis of inexact accelerated minibatch SVRG

The main results established in Section 3 relies on the analysis for inexact, minibatch, accelerated SVRG update (IMBA-SVRG Algorithm 3), where in each stochastic step, we allow a small error ξ_t in the updating. In this section, we show that as long as the inexactness at each iteration is small enough, IMBA-SVRG can use a large minibatch size without slowing down the convergence. We first state the main theorem which characterize the rate of convergence, then explain the prove mechanism.

Algorithm 3 IMBA-SVRG: Inexact Minibatch Accelerated SVRG Method.

Parameters $\alpha = \sqrt{\eta\lambda/2}$.

Initialize $\tilde{w}_0 = 0$.

for $s = 1, 2, \dots$ **do**

Calculate $\tilde{v} = \frac{1}{n} \sum_{i=1}^n \nabla f_i(\tilde{w}_{s-1})$.

Initialize $y_0 = w_0 = \tilde{w}_s$.

for $t = 1, 2, \dots, m$ **do**

Sampling b items from $[n]$ to form a minibatch B_t .

Inexact update w_t such that

$$w_t = \bar{w}_t + \xi_t, \quad (18)$$

where:

$$\bar{w}_t = y_{t-1} - \eta \left(\frac{1}{b} \sum_{i \in B_t} \nabla f_i(y_{t-1}) - \frac{1}{b} \sum_{i \in B_t} \nabla f_i(\tilde{w}_{s-1}) + \tilde{v} \right).$$

Update

$$y_t = w_t + \left(\frac{1 - \alpha}{1 + \alpha} \right) (w_t - w_{t-1}).$$

end for

Update $\tilde{w}_s = w_m$.

end for

Return \tilde{w}_s

Theorem 12. For IMBA-SVRG algorithm, If we choose stepsize as

$$\eta = \min \left\{ \frac{b^2 \lambda}{6400L^2}, \frac{1}{8L} \right\},$$

and when the deviation ξ_k satisfies $\forall 1 \leq k \leq t$

$$\|\xi_k\|^2 \leq \frac{2\lambda^2 \eta^3 \sqrt{\lambda\eta}}{15} (f(\tilde{w}_{s-1}) - f(w^*)),$$

and the number of iterations t satisfying

$$t \geq \frac{10}{9\sqrt{\lambda\eta}} \log(36),$$

then

$$\mathbb{E}[f(\tilde{w}_s) - f(w^*)] \leq \frac{1}{2} [f(\tilde{w}_{s-1}) - f(w^*)].$$

Remark 4. Depending on the relative magnitude of minibatch size b and condition number $\frac{L}{\lambda}$, the iteration complexity can be interpreted as two phases: (i) when the minibatch size b satisfies $b \leq 20\sqrt{\frac{2L}{\lambda}}$, then if we choose stepsize as $\eta = \frac{b^2 \lambda}{6400L^2}$, and the deviation ξ_k satisfies $\forall 1 \leq k \leq t$: $\|\xi_k\|^2 \leq \frac{2\lambda^2 \eta^3 \sqrt{\lambda\eta}}{15} (f(\tilde{w}_{s-1}) - f(w^*))$, we know when the iteration number satisfies $t \geq \frac{90L}{\lambda b}$, then $\mathbb{E}[f(\tilde{w}_s) - f(w^*)] \leq \frac{1}{2} [f(\tilde{w}_{s-1}) - f(w^*)]$; (ii) when the minibatch size satisfies $20\sqrt{\frac{2L}{\lambda}} \leq b \leq n$,

then if we choose stepsize as $\eta = \frac{1}{8L}$, and when the deviation ξ_k satisfies $\forall 1 \leq k \leq t: \|\xi_k\|^2 \leq \frac{2\lambda^2\eta^3\sqrt{\lambda\eta}}{15}(f(\tilde{w}_{s-1}) - f(w^*))$, we know when $t \geq \sqrt{\frac{8L}{\lambda}}$, then $\mathbb{E}[f(\tilde{w}_s) - f(w^*)] \leq \frac{1}{2}[f(\tilde{w}_{s-1}) - f(w^*)]$.

A direct consequence of Theorem 12 is the following iteration complexity of IMBA-SVRG.

Corollary 13. *For IMBA-SVRG algorithm, If we choose stepsize as $\eta = \min\left\{\frac{b^2\lambda}{6400L^2}, \frac{1}{8L}\right\}$, and when at every state s , the deviation ξ_k at every iteration satisfies $\forall 1 \leq k \leq t: \|\xi_k\|^2 \leq \frac{2\lambda^2\eta^3\sqrt{\lambda\eta}}{15}(f(\tilde{w}_{s-1}) - f(w^*))$, then $\mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$ full gradient evaluations and*

$$\mathcal{O}\left(\max\left\{\frac{L}{\lambda b}, \sqrt{\frac{L}{\lambda}}\right\} \cdot \log\left(\frac{1}{\epsilon}\right)\right)$$

inexact update steps of form (18) are sufficient to ensure $\mathbb{E}f(\tilde{w}_s) - f(w^) \leq \epsilon$.*

Our proof of convergence for IMBA-SVRG relies on the machinery of stochastic estimation sequence (Lin et al., 2014), which originated from the framework of estimation sequence developed in (Nesterov, 2004). The main difference with (Lin et al., 2014) is here we no longer require the inequalities in estimation sequences hold almost surely, rather we only make sure they hold in expectation, thus the quadratic lower bound used to construct the stochastic estimation sequence is also different. IMBA-SVRG can be viewed as an inexact extension of accelerated minibatch SVRG (Nitanda, 2014), though (Nitanda, 2014) also considered stochastic estimation sequence, here we allow error in the stochastic gradient update thus construct a different estimation sequence which takes the inexactness into consideration; on the other hand, batch (accelerated) gradient methods with inexact first order oracle have been studied in (Schmidt et al., 2011; Villa et al., 2013; Devolder et al., 2014), the analysis of IMBA-SVRG extends these results in the context of stochastic gradient methods with variance reduction.

We started by giving the definition of stochastic estimation sequence.

Definition 14. (Stochastic estimation sequence) *A sequence of pairs $\{V_t(w), \theta_t\}_{t \geq 0}$ is called an estimation sequence of the function $f(w)$ if $\theta_t > 0$ and for any $w \in \mathbb{R}^d, t \geq 0$ we have*

$$V_t(w) \leq (1 - \theta_t)f(w) + \theta_t V_0(w), \quad (19)$$

if $\{V_t(w)\}_{t \geq 0}$ is a sequence of random functions, we call the sequence of pairs $\{V_t(w), \theta_t\}_{t \geq 0}$ a stochastic estimation sequence if (19) holds in expectation, i.e.

$$\mathbb{E}[V_t(w)] \leq (1 - \theta_t)f(w) + \theta_t V_0(w).$$

The following lemma is a generalized version of Lemma 1 in (Lin et al., 2014), which shows if we can construct upper bound of $\mathbb{E}[f(w_t)]$ using $\mathbb{E}[V_t(w)]$, then we get convergence of $\mathbb{E}[f(w_t)]$.

Lemma 15. *Suppose $\{V_t(w), \theta_t\}_{t \geq 0}$ is a stochastic estimation sequence of the function $f(w)$. Let w^* be the minimizer of $f(w)$. If there are sequences of random variables $\{w_t\}_{t \geq 0}$ and $\{\epsilon_t\}_{t \geq 0}$ in \mathbb{R}^d , $\{\delta_t\}_{t \geq 0}$ in \mathbb{R} such that*

$$\mathbb{E}[f(w_t)] \leq \min_w \{\mathbb{E}[V_t(w)]\} + \mathbb{E}[\delta_t] \quad (20)$$

holds for all $t \geq 0$, then

$$\mathbb{E}[f(w_t)] - f(w^*) \leq \theta_t(V_0(w^*) - f(w^*)) + \mathbb{E}[\delta_t].$$

The following lemma construct a concrete stochastic estimation sequence.

Lemma 16. Assume $f(w) = \frac{1}{n} \sum_{i=1}^n f_i(w)$, where each $f_i(w)$ is λ -strongly convex and L -smooth. Suppose that

- $V_0(w)$ is an arbitrary deterministic function on \mathbb{R}^d ;
- $\{\alpha_t\}_{t \geq 0}$ is a sequence that satisfies $\alpha_t \in (0, 1), \forall t \geq 0$, and $\sum_{t=0}^{\infty} \alpha_t = \infty$.
- $\{y_t\}_{t=0}^{\infty}$ is an arbitrary sequence in \mathbb{R}^d ;
- Define $\{v_t\}_{t \geq 1}$ as:

$$v_t = \frac{1}{b} \sum_{i \in B_t} \nabla f_i(y_{t-1}) - \frac{1}{b} \sum_{i \in B_t} \nabla f_i(\tilde{w}_{s-1}) + \frac{1}{n} \sum_{i=1}^n \nabla f_i(\tilde{w}_{s-1}).$$

Define the sequence $\{w_t\}_{t \geq 0}$ and $\{V_t(w)\}_{t \geq 0}$ as follows. Let $y_0 = w_0$ and w_t be arbitrary vector in \mathbb{R}^d such that

$$w_t = y_{t-1} - \eta v_t + \xi_t,$$

where the stepsize η satisfies $0 < \eta \leq \frac{1}{L}$. Let

$$\begin{aligned} V_t(w) = & (1 - \alpha_{t-1})V_{t-1}(w) \\ & + \alpha_{t-1} \left(\frac{1}{b} \sum_{i \in B_t} f_i(y_{t-1}) + \left\langle v_t - \frac{\xi_t}{\eta}, w - y_{t-1} \right\rangle + \frac{\lambda}{4} \|w - y_{t-1}\|^2 - \frac{\|\xi_t\|^2}{\lambda \eta^2} \right), \end{aligned} \quad (21)$$

let $\theta_0 = 1$, and $\theta_t = (1 - \alpha_{t-1})\theta_{t-1}, \forall t \geq 1$. Then the sequence $\{V_t(w), \theta_t\}_{t \geq 0}$ is a stochastic estimation sequence of $f(w)$.

By above construction, we know $V_t(w)$ is a quadratic function, its minimizer and minimum value has the following iterative form:

Lemma 17. Define $\forall t \geq 0$:

$$V_t^* = \min_w V_t(w),$$

suppose we choose the function $V_0(w)$ in Lemma 16 as

$$V_0(w) = V_0^* + \frac{\lambda}{4} \|w - z_0\|^2,$$

with $V_0^* = f(z_0), \epsilon_0 = 0$ and $\delta_0 = 0$ for $z_0 = w_0$. Then the sequence $\{V_t(w)\}_{t \geq 0}$ defined in 21 can be written as

$$V_t(w) = V_t^* + \frac{\lambda}{4} \|w - z_t\|^2, \quad (22)$$

where the sequences $\{V_t^*\}, \{z_t\}$ are defined recursively as:

$$z_t = (1 - \alpha_{t-1})z_{t-1} + \alpha_{t-1}y_{t-1} - \frac{2\alpha_{t-1}}{\lambda} \left(v_t - \frac{\xi_t}{\eta} \right), \quad (23)$$

$$\begin{aligned} V_t^* = & (1 - \alpha_{t-1})V_{t-1}^* + \frac{\alpha_{t-1}(1 - \alpha_{t-1})\lambda}{4} \|z_{t-1} - y_{t-1}\|^2 + \alpha_{t-1}(1 - \alpha_{t-1}) \left\langle v_t - \frac{\xi_t}{\eta}, z_{t-1} - y_{t-1} \right\rangle \\ & - \frac{\alpha_{t-1}^2}{\lambda} \|v_t\|^2 + \frac{\alpha_{t-1}}{b} \sum_{i \in B_t} f_i(y_{t-1}) - \frac{(\alpha_{t-1} + \alpha_{t-1}^2) \|\xi_t\|^2}{\lambda \eta^2} + \frac{2\alpha_{t-1}^2}{\lambda \eta} \langle \xi_t, v_t \rangle. \end{aligned} \quad (24)$$

The following lemma establishes a connection between z_t, y_t, w_t when z_t is updated as (23), and y_t, w_t is updated as Algorithm 3.

Lemma 18. *Based on the updating rule of z_t in (23), if we choose $\alpha_t = \alpha = \sqrt{\frac{\lambda\eta}{2}}, \forall t \geq 0$, we have the following inequalities hold for all $t \geq 0$:*

$$z_t - y_t = \frac{1}{\alpha}(y_t - w_t).$$

The following lemma gives concrete expression of δ_t such that for the IMBA-SVRG algorithm, the condition (20) in Lemma 15 is satisfied.

Lemma 19. *Suppose we choose $\forall t \geq 0$*

$$\alpha_t = \alpha = \sqrt{\frac{\lambda\eta}{2}},$$

and the stepsize η satisfying $\eta \leq \frac{1}{8L}$, we have $\forall t \geq 0$

$$\mathbb{E}[f(w_t)] \leq \mathbb{E}[V_t^*] + \mathbb{E}[\delta_t],$$

where $\delta_0 = 0$ and $\forall t \geq 1$:

$$\delta_t = \left(1 - \sqrt{\frac{\lambda\eta}{2}}\right) \delta_{t-1} + \left(\eta \|v_t - \nabla f(y_{t-1})\|^2 - \frac{(1 - \sqrt{\lambda\eta/2})\lambda}{\sqrt{2\lambda\eta}} \|y_{t-1} - w_{t-1}\|^2 + \frac{1}{4\lambda^2\eta^3} \|\xi_t\|^2\right).$$

The lemma below bounds the term $\mathbb{E}\left[\eta \|v_t - \nabla f(y_{t-1})\|^2 - \frac{(1 - \sqrt{\lambda\eta/2})\lambda}{\sqrt{2\lambda\eta}} \|y_{t-1} - w_{t-1}\|^2\right]$.

Lemma 20. *Suppose B_t is constructed by uniform sampling with or without replacement, we have the following inequality holds:*

$$\begin{aligned} \mathbb{E}\left[\eta \|v_t - \nabla f(y_{t-1})\|^2 - \frac{(1 - \sqrt{\lambda\eta/2})\lambda}{\sqrt{2\lambda\eta}} \|y_{t-1} - w_{t-1}\|^2\right] &\leq \frac{8\eta L}{b} (f(w_{t-1}) - f(w^*) + f(\tilde{w}_{s-1}) - f(w^*)) \\ &\quad + \left(\frac{2\eta L^2}{b} - \frac{(1 - \sqrt{\lambda\eta/2})\lambda}{\sqrt{2\lambda\eta}}\right) \|y_{t-1} - w_{t-1}\|^2. \end{aligned}$$

Based on above lemma, by carefully choosing the stepsize η , we get the following lemma which is important to obtain iteration complexity for IMBA-SVRG.

Lemma 21. *If we choose the stepsize satisfying*

$$\eta \leq \min\left\{\frac{b^2\lambda}{6400L^2}, \frac{1}{8L}\right\}$$

then the following inequality holds $\forall t \geq 0$:

$$\begin{aligned} \mathbb{E}f(w_t) - f(w^*) &\leq \left(1 - \sqrt{\frac{\lambda\eta}{2}}\right)^t (V_0(w^*) - f(w^*)) \\ &\quad + \mathbb{E}\left[\sum_{k=1}^t \left(1 - \sqrt{\frac{\lambda\eta}{2}}\right)^{t-k} \left(\frac{1}{4\lambda^2\eta^3} \|\xi_k\|^2 + \frac{8\eta L}{b} (f(w_{k-1}) + f(\tilde{w}_{s-1}) - 2f(w^*))\right)\right], \end{aligned}$$

Table 2: List of datasets used in the experiments.

Name	#Instances	#Features	Task
codrna	59,535	8	Classification
covtype	581,012	54	Classification
svmguide3	1,243	21	Classification
synthetic-c	10,000	1,000	Classification
cadata	20,460	8	Regression
spacega	3,107	6	Regression
synthetic-r	20,000	2,000	Regression
year	463,715	91	Regression

5 Experiments

In this section we compare the proposed MP-SVRP algorithm to several state-of-the-art methods in minimizing (1). The used datasets are summarized in Table 2, most of which can be download from the LibSVM website³. We also considered a synthetic dataset for each task where the data $\{x_i, b_i\}_{i=1}^n$ are i.i.d. drawn from the following model:

$$\text{Regression : } b_i = \langle x_i, \bar{w} \rangle + a_i, \quad x_i \sim \mathcal{N}(0, \Sigma), \quad a_i \sim \mathcal{N}(0, 1), \quad \forall i \in [n],$$

$$\text{Classification : } P(b_i = \pm 1) = \frac{\exp(b_i \langle x_i, \bar{w} \rangle)}{1 + \exp(b_i \langle x_i, \bar{w} \rangle)}, \quad x_i \sim \mathcal{N}(0, \Sigma), \quad \forall i \in [n],$$

where entries of \bar{w} are drawn i.i.d. from $\mathcal{N}(0, 1)$. To make the problem ill-conditioned with fast decaying spectrum, we set $\Sigma_{ij} = 2^{-|i-j|/500}$, $\forall i, j \in [n]$. We consider ridge regression and logistic regression models for regression and classification problems, respectively. We normalize the dataset by $x_i \leftarrow x_i / (\max_i \|x_i\|^2)$ to ensure the maximum norm of data points is 1, this makes the setting regularization parameter λ easier. We tried three settings of λ , as $1/n$, $10^{-1}/n$ and $10^{-2}/n$ to represent different levels of regularization. It is expected that when $\lambda = 1/n$, the SVRG/SAGA/SDCA algorithms should converge very fast since the condition number is of the same order as sample size n , while for weak regularization case $\lambda = 10^{-2}/n$ these algorithms are expected to converge slow.

We compare with SVRG (Johnson and Zhang, 2013), SDCA (Shalev-Shwartz and Zhang, 2013), SAGA Defazio et al. (2014) which represent popular variance reduced optimization algorithms, we also compare with a related minibatch accelerated SVRG (MB-SVRG) (Nitanda, 2014) which allows large minibatch size without slowing down convergence. For quasi-Newton methods, we compare with L-BFGS with 10 as memory size, and with line search of stepsize to satisfy the Wolfe condition (Wright and Nocedal, 1999). For the proposed MB-SVRP method, at every iteration we simply run one pass of SVRG initialized with y_{t-1} , to approximately solve (7). Other parameters, such as stepsize in SVRG and SAGA, the minibatch size of MP-SVRG, are tuned to give the fastest convergence, the minibatch size of MB-SVRG is set to be the same as MB-SVRP to demonstrate the direct comparison.

Figure 2 and 3 showed the results for ℓ_2 regularized logistic regression and ridge regression, respectively, where we plot how the objective suboptimality $f(w_t) - f(w^*)$ decrease as the number of gradient evaluations divided by sample size (a.k.a. number of effective passes) increase. We have the following observations:

³<https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/>

- When the regularization parameter λ (thus strong convexity) is large enough, all methods (except L-BFGS) converges very fast, typically using less than 50 passes over data to converges to numerical precision.
- L-BFGS typically converges the slowest, especially when λ is large, the advantages of variance-reduced stochastic methods over L-BFGS are significant.
- When the regularization parameter is small, the proposed MP-SVRP method started to show advantages, especially when $\lambda = 10^{-2}/n$, MP-SVRP is substantially much faster than all other methods compared.

5.1 Empirical results for composite optimization

We also consider empirical comparisons in the setting of non-smooth composition optimization problems. We adopted the same datasets used for the smooth optimization problems, but considered the elastic-net regularized logistic regression and linear regression models (Zou and Hastie, 2005). Where the objective is in the form of (10) but with a elastic-net regularization:

$$g(w) = \frac{\lambda}{2} \|w\|^2 + \mu \|w\|_1,$$

where μ is set to be $10^{-1}/n$. We follow the same setting as previous section, and compared with prox SVRG (Xiao and Zhang, 2014), prox SAGA (Defazio et al., 2014) and prox MB-SVRG (Nitanda, 2014), but didn't compare with L-BFGS because it is generally only applicable for smooth optimization problems. The results are shown in Figure 4 and 5, we have pretty much similar observations as the smooth optimization case: all variance-reduced methods converges reasonably fast when λ is large, but when λ becomes small, the proposed prox MP-SVRP method converges significantly faster because it effectively leveraged the higher-order information from minibatches.

6 Conclusion

In this paper, we propose an novel minibatch stochastic optimization approach for regularized loss minimization in machine learning, it efficiently utilize both variance-reduced gradients and sub-sampled higher order information, it provably improves over the iteration complexity of previous state-of-the-art in certain scenarios, and performs well on a variety of smooth and composite optimization tasks in practice. The minibatch nature of the algorithm makes it has potential to be implemented under parallel and distributed computing environment, where additional speed up can be obtained with more computing resources.

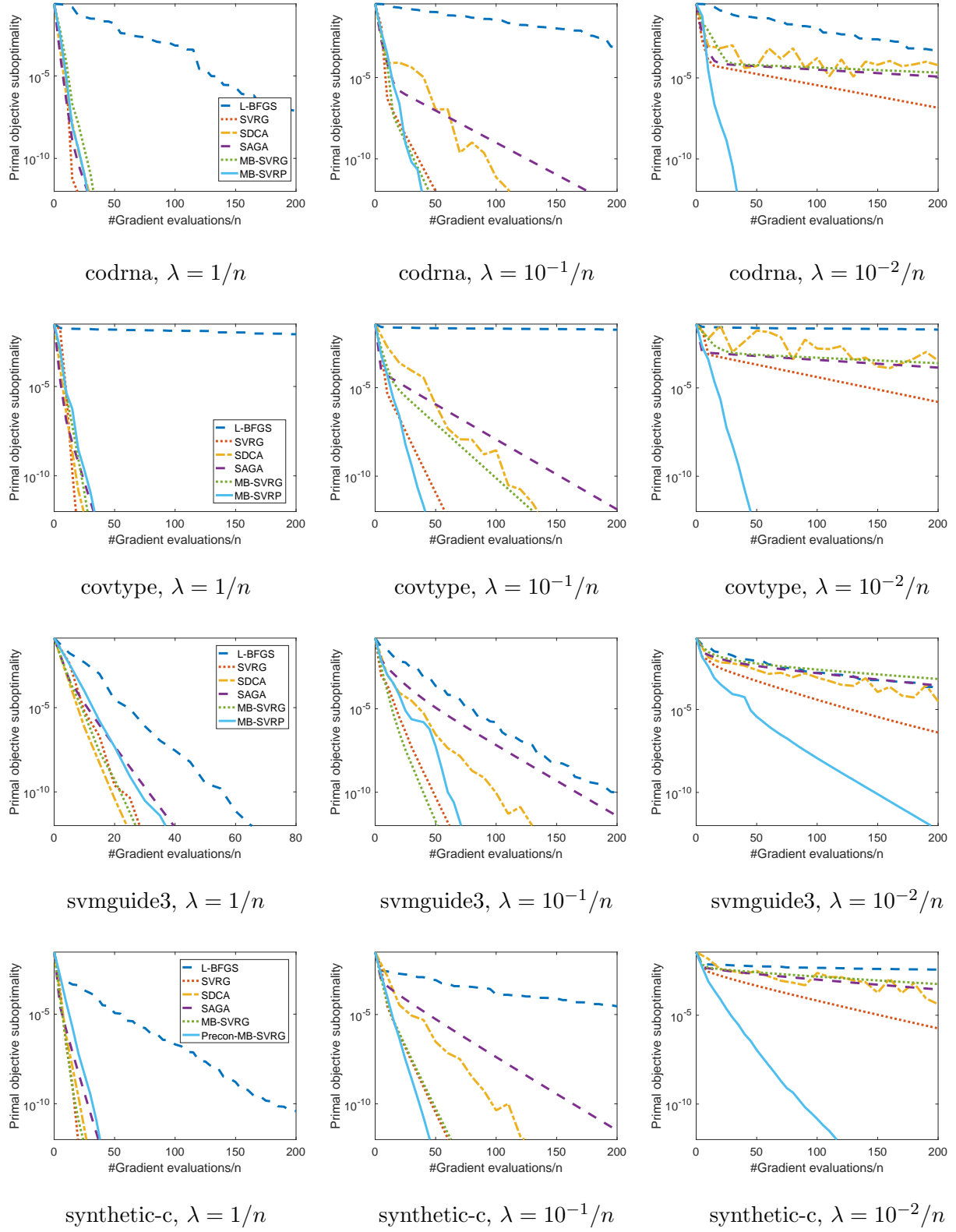


Figure 2: Comparison of various optimization algorithms for solving ℓ_2 regularized logistic regression problems.

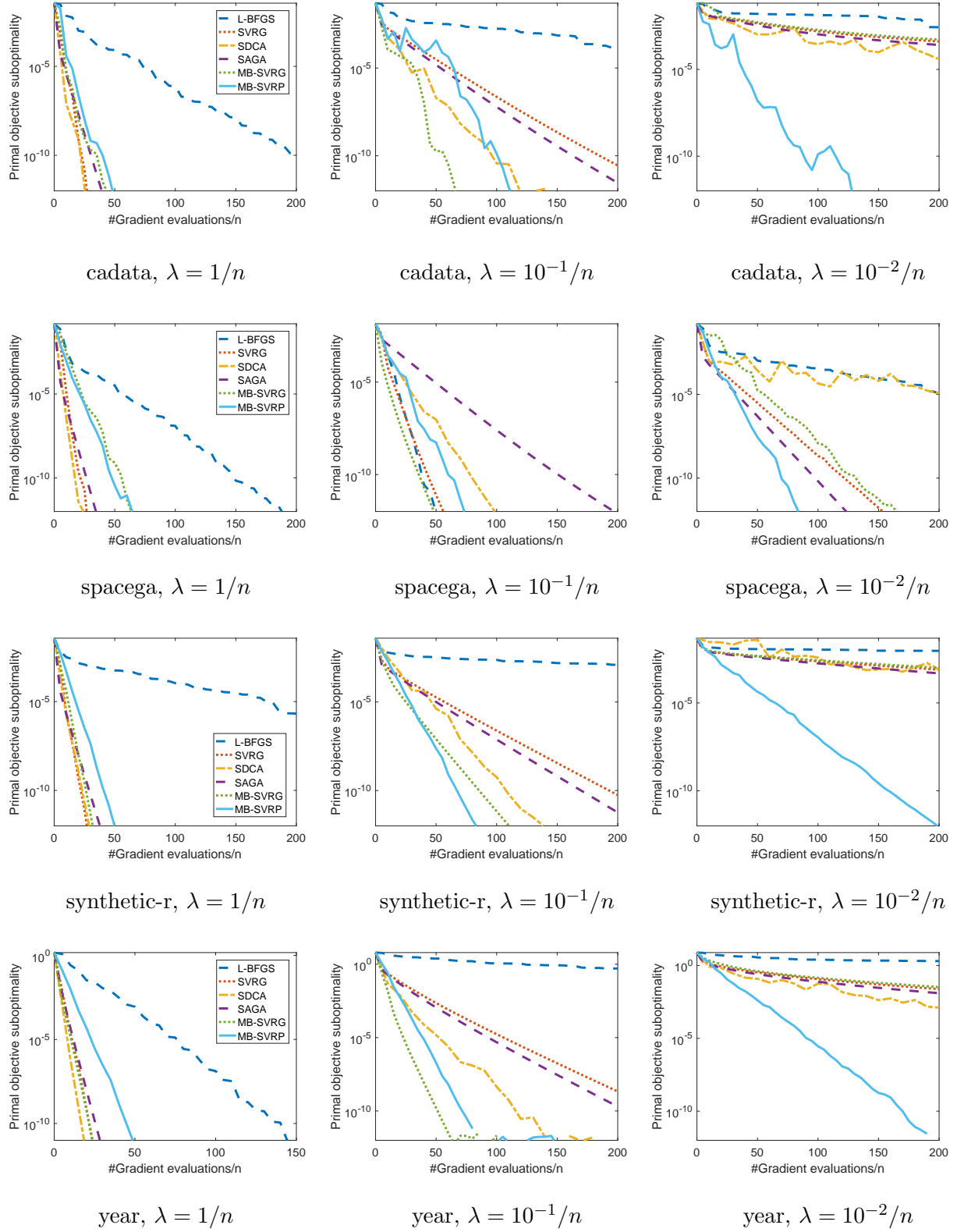


Figure 3: Comparison of various optimization algorithms for solving ridge regression problems.

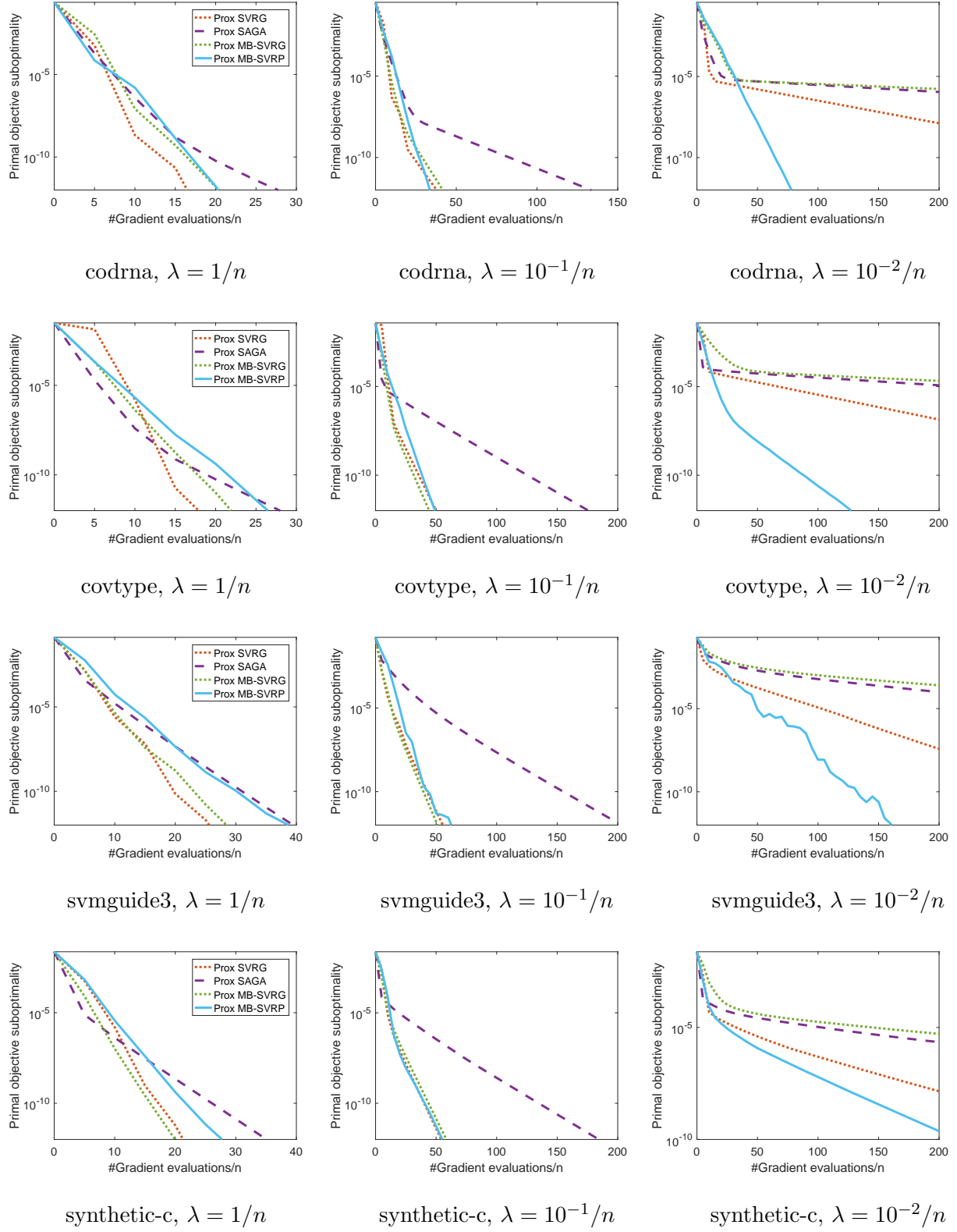


Figure 4: Comparison of various optimization algorithms for solving elastic-net regularized logistic regression problems.

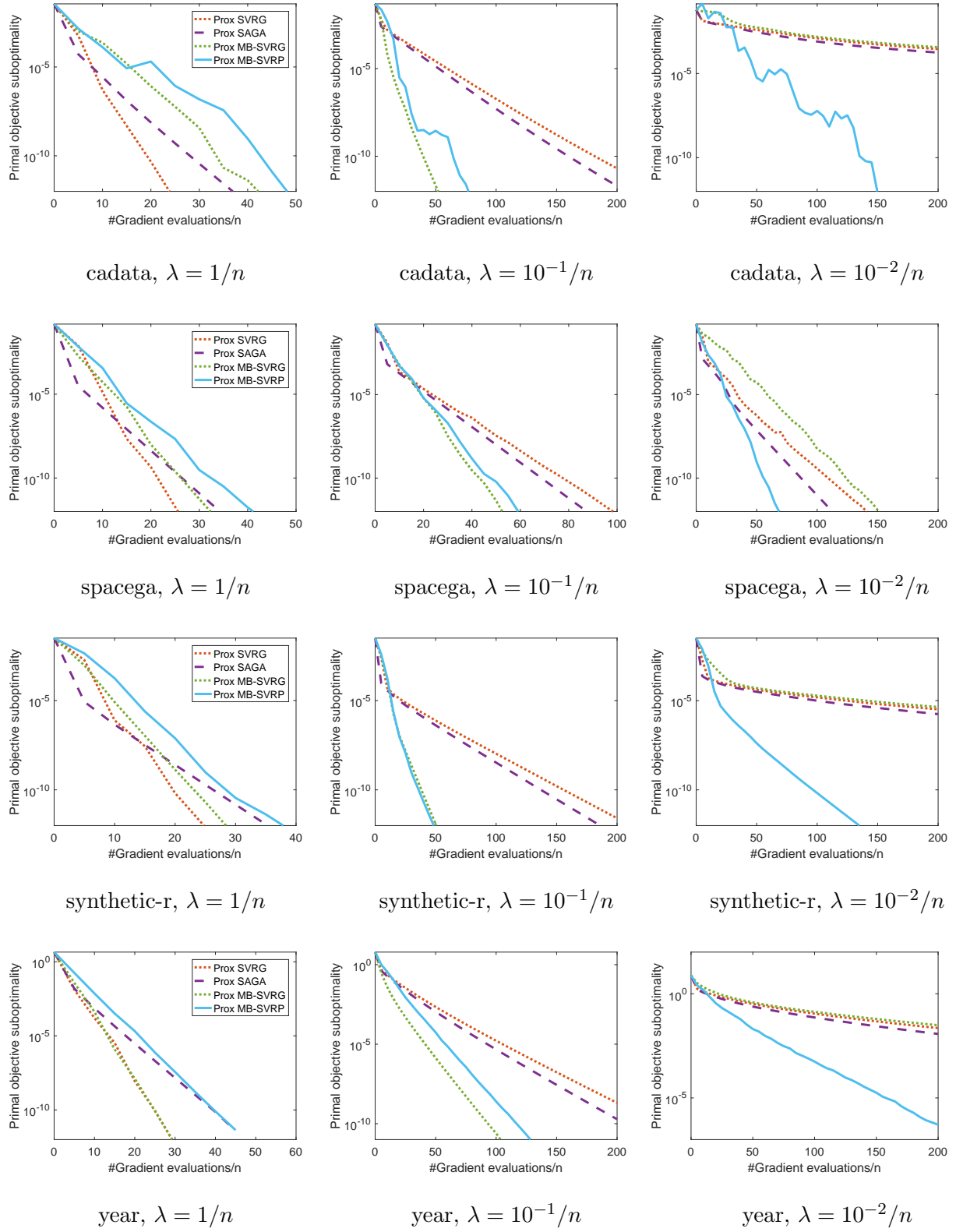


Figure 5: Comparison of various optimization algorithms for solving elastic-net regularized regression problems.

Appendix

The appendix contains proofs of some theorems and lemmas stated in the main paper.

A Proofs in Section 3

A.1 Proof of Proposition 3

Proof. Let $w^* = \arg \min_w f(w)$, and $z^* = \arg \min_z f(\tilde{H}_\lambda^{-1/2} z)$, it is clear that $w^* = \tilde{H}_\lambda^{-1/2} z^*$, which is coincident with the exchange of variable rule $z = \tilde{H}_\lambda^{1/2} w$. Moreover, by (14) we have

$$\begin{aligned} z_t &\leftarrow z_{t-1} - \eta \left(\frac{1}{b} \sum_{i \in B_t} \nabla_z f_i(\tilde{H}_\lambda^{-1/2} z_{t-1}) + \frac{1}{n} \sum_{i=1}^n \nabla_z f_i(\tilde{H}_\lambda^{-1/2} \tilde{z}) - \frac{1}{b} \sum_{i \in B_t} \nabla_z f_i(\tilde{H}_\lambda^{-1/2} \tilde{z}) \right) \\ &\stackrel{\textcircled{1}}{=} z_{t-1} - \eta H_\lambda^{-1/2} \left(\frac{1}{b} \sum_{i \in B_t} \nabla f_i(\tilde{H}_\lambda^{-1/2} z_{t-1}) + \frac{1}{n} \sum_{i=1}^n \nabla f_i(\tilde{H}_\lambda^{-1/2} \tilde{z}) - \frac{1}{b} \sum_{i \in B_t} \nabla f_i(\tilde{H}_\lambda^{-1/2} \tilde{z}) \right) \\ &\stackrel{\textcircled{2}}{=} z_{t-1} - \eta H_\lambda^{-1/2} \left(\frac{1}{b} \sum_{i \in B_t} \nabla f_i(w_{t-1}) + \frac{1}{n} \sum_{i=1}^n \nabla f_i(\tilde{w}) - \frac{1}{b} \sum_{i \in B_t} \nabla f_i(\tilde{w}) \right), \end{aligned}$$

where at step ① we have used the gradient chain rule, at step ② we used the exchange of variables $z = \tilde{H}_\lambda^{1/2} w$. Multiplying both sides of above equation with $\tilde{H}_\lambda^{-1/2}$ we recover the exact formulation of (13). \square

A.2 Proof of Lemma 4

Proof. Based on $\lambda + \tilde{\lambda}$ -strong convexity of $\tilde{f}_t(w)$, we have

$$\frac{\lambda + \tilde{\lambda}}{2} \|w_t - \bar{w}_t\|^2 \leq \tilde{f}_t(w_t) - \tilde{f}_t(\bar{w}_t) = \varepsilon,$$

thus

$$\|z_t - \bar{z}_t\| = \left\| \tilde{H}_\lambda^{1/2} (w_t - \bar{w}_t) \right\| \leq \left\| \tilde{H}_\lambda \right\|^{1/2} \|w_t - \bar{w}_t\| \leq \sqrt{L + \tilde{\lambda}} \|w_t - \bar{w}_t\| \leq \sqrt{\frac{2\varepsilon(L + \tilde{\lambda})}{\lambda + \tilde{\lambda}}}.$$

\square

A.3 Proof of Lemma 7

Proof. For $j = 1, \dots, b$, define random matrix

$$v_j = \frac{1}{b} \left(H_\lambda^{-1/2} x_k x_k^\top H_\lambda^{-1/2} - \frac{1}{n} \sum_{i=1}^n H_\lambda^{-1/2} x_i x_i^\top H_\lambda^{-1/2} \right),$$

with probability $1/n, \forall k \in [n]$. It is easy to check that $\mathbb{E}[v_j] = 0$, and

$$\begin{aligned} \|v_j\| &\leq \frac{1}{b} \left(\left\| H_{\tilde{\lambda}}^{-1/2} x_k x_k^\top H_{\tilde{\lambda}}^{-1/2} \right\| + \left\| \frac{1}{n} \sum_{i=1}^n H_{\tilde{\lambda}}^{-1/2} x_i x_i^\top H_{\tilde{\lambda}}^{-1/2} \right\| \right) \\ &\leq \frac{1}{b} \left(2 \max_{i \in [n]} x_i^\top H_{\tilde{\lambda}}^{-1} x_i \right) \leq \frac{2\rho_{\tilde{\lambda}}^2}{b} \left(\frac{1}{n} \sum_{i=1}^n x_i^\top H_{\tilde{\lambda}}^{-1} x_i \right) \\ &= \frac{2\rho_{\tilde{\lambda}}^2}{b} \text{tr} \left(H_{\tilde{\lambda}}^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right) \right) = \frac{2\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}}}{b}. \end{aligned}$$

for the expected second order moment, denote $\tilde{x}_k = H_{\tilde{\lambda}}^{-1/2} x_k$ we have

$$\begin{aligned} \|\mathbb{E}[v_j^2]\| &= \left\| \frac{1}{nb^2} \sum_{k=1}^n \left(\tilde{x}_k \tilde{x}_k^\top - \frac{1}{n} \sum_{i=1}^n \tilde{x}_i \tilde{x}_i^\top \right)^2 \right\| \\ &= \left\| \frac{1}{nb^2} \sum_{k=1}^n \left(\left(\tilde{x}_k \tilde{x}_k^\top \right)^2 - 2 \left(\tilde{x}_k \tilde{x}_k^\top \right) \left(\frac{1}{n} \sum_{i=1}^n \tilde{x}_i \tilde{x}_i^\top \right) + \left(\frac{1}{n} \sum_{i=1}^n \tilde{x}_i \tilde{x}_i^\top \right)^2 \right) \right\| \\ &= \left\| \frac{1}{nb^2} \sum_{k=1}^n \left(\tilde{x}_k \tilde{x}_k^\top \right)^2 - \frac{2}{b^2} \left(\frac{1}{n} \sum_{i=1}^n \tilde{x}_i \tilde{x}_i^\top \right)^2 + \frac{1}{b^2} \left(\frac{1}{n} \sum_{i=1}^n \tilde{x}_i \tilde{x}_i^\top \right)^2 \right\| \\ &= \left\| \frac{1}{nb^2} \sum_{k=1}^n \left(\tilde{x}_k \tilde{x}_k^\top \right)^2 - \frac{1}{b^2} \left(\frac{1}{n} \sum_{i=1}^n \tilde{x}_i \tilde{x}_i^\top \right)^2 \right\| \\ &\leq \left\| \frac{1}{nb^2} \sum_{k=1}^n \left(\tilde{x}_k \tilde{x}_k^\top \right)^2 \right\| \leq \left\| \frac{\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}}}{b^2} \left(\frac{1}{n} \sum_{k=1}^n \tilde{x}_k \tilde{x}_k^\top \right) \right\| = \frac{\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}}}{b^2} \left\| H_{\tilde{\lambda}}^{-1/2} \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right) H_{\tilde{\lambda}}^{-1/2} \right\| \\ &\leq \frac{\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}}}{b^2}. \end{aligned}$$

Thus

$$\left\| \sum_{j=1}^b \mathbb{E}[v_j^2] \right\| \leq \frac{\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}}}{b},$$

using Lemma 22, we know

$$P \left(\left\| \frac{1}{b} \sum_{i \in \tilde{B}} x_i x_i^\top - \frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right\|_2 \geq t \right) = P \left(\left\| \sum_{j=1}^b v_j \right\| \geq t \right) \leq d \exp \left(\frac{-t^2/2}{\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}}/b + 2\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}} t/(3b)} \right),$$

setting $t = 2 \log \left(\frac{d}{\delta} \right) \cdot \sqrt{\frac{\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}}}{b}}$ we conclude the proof. \square

A.4 Proof of Lemma 8

Proof. When $\tilde{\lambda} \geq \lambda$, we know

$$\begin{aligned}\lambda_{\min}(\tilde{H}_{\tilde{\lambda}}^{-1/2} H_{\lambda} \tilde{H}_{\tilde{\lambda}}^{-1/2}) &\geq \lambda_{\min} \left(\tilde{H}_{\tilde{\lambda}}^{-1/2} \left(\frac{\lambda}{\tilde{\lambda}} \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^{\top} \right) + \lambda I \right) \tilde{H}_{\tilde{\lambda}}^{-1/2} \right) \\ &= \frac{\lambda}{\tilde{\lambda}} \lambda_{\min}(\tilde{H}_{\tilde{\lambda}}^{-1/2} H_{\tilde{\lambda}} \tilde{H}_{\tilde{\lambda}}^{-1/2}).\end{aligned}$$

Further more we can lower bound $\lambda_{\min}(\tilde{H}_{\tilde{\lambda}}^{-1/2} H_{\tilde{\lambda}} \tilde{H}_{\tilde{\lambda}}^{-1/2})$ by

$$\begin{aligned}\lambda_{\min}(\tilde{H}_{\tilde{\lambda}}^{-1/2} H_{\tilde{\lambda}} \tilde{H}_{\tilde{\lambda}}^{-1/2}) &= \lambda_{\min}(\tilde{H}_{\tilde{\lambda}}^{-1} H_{\tilde{\lambda}}) \\ &= \frac{1}{\lambda_{\max}(H_{\tilde{\lambda}}^{-1} \tilde{H}_{\tilde{\lambda}})} \\ &\geq \frac{1}{\lambda_{\max}(H_{\tilde{\lambda}}^{-1} H_{\tilde{\lambda}}) + \lambda_{\max}(H_{\tilde{\lambda}}^{-1} (H_{\tilde{\lambda}} - \tilde{H}_{\tilde{\lambda}}))} \\ &\geq \frac{1}{1 + 2 \log \left(\frac{d}{\delta} \right) \sqrt{\frac{\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}}}{b}}},\end{aligned}$$

where the last step we used Lemma (7). □

A.5 Proof of Lemma 9

Proof. First we upper bound $\lambda_{\max}(\tilde{H}_{\tilde{\lambda}}^{-1/2} H_{\tilde{\lambda}} \tilde{H}_{\tilde{\lambda}}^{-1/2})$, since when $b \geq 16\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}} \log^2(d/\delta)$ we have $\lambda_{\max}(H_{\tilde{\lambda}}^{-1} (H_{\tilde{\lambda}} - \tilde{H}_{\tilde{\lambda}})) \leq 1$, thus

$$\begin{aligned}\lambda_{\max}(\tilde{H}_{\tilde{\lambda}}^{-1/2} H_{\tilde{\lambda}} \tilde{H}_{\tilde{\lambda}}^{-1/2}) &= \frac{1}{\lambda_{\min}(H_{\tilde{\lambda}}^{-1} \tilde{H}_{\tilde{\lambda}})} \\ &\leq \frac{1}{1 - \lambda_{\max}(H_{\tilde{\lambda}}^{-1} (H_{\tilde{\lambda}} - \tilde{H}_{\tilde{\lambda}}))} \\ &\leq \frac{1}{1 - 2 \log \left(\frac{d}{\delta} \right) \sqrt{\frac{\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}}}{b}}} \\ &\leq 2.\end{aligned}$$

Then we bound $\max_{i \in [n]} x_i^{\top} \tilde{H}_{\tilde{\lambda}}^{-1} x_i^{\top}$ is through $x_i^{\top} H_{\tilde{\lambda}}^{-1} x_i^{\top}$, because

$$\begin{aligned}x_i^{\top} \tilde{H}_{\tilde{\lambda}}^{-1} x_i^{\top} &= x_i^{\top} H_{\tilde{\lambda}}^{-1} x_i^{\top} + x_i^{\top} (\tilde{H}_{\tilde{\lambda}}^{-1} - H_{\tilde{\lambda}}^{-1}) x_i^{\top} \\ &= x_i^{\top} H_{\tilde{\lambda}}^{-1} x_i^{\top} + x_i^{\top} (\tilde{H}_{\tilde{\lambda}}^{-1} H_{\tilde{\lambda}} - I) H_{\tilde{\lambda}}^{-1} x_i^{\top} \\ &\leq x_i^{\top} H_{\tilde{\lambda}}^{-1} x_i^{\top} + \lambda_{\max}(\tilde{H}_{\tilde{\lambda}}^{-1} H_{\tilde{\lambda}} - I) x_i^{\top} H_{\tilde{\lambda}}^{-1} x_i^{\top} \\ &\stackrel{\textcircled{1}}{\leq} 2 x_i^{\top} H_{\tilde{\lambda}}^{-1} x_i^{\top},\end{aligned}$$

where in step ① we used the fact that

$$\lambda_{\max}(\tilde{H}_{\tilde{\lambda}}^{-1} H_{\tilde{\lambda}} - I) \leq \max \left\{ |\lambda_{\max}(\tilde{H}_{\tilde{\lambda}}^{-1} H_{\tilde{\lambda}}) - 1|, |\lambda_{\min}(\tilde{H}_{\tilde{\lambda}}^{-1} H_{\tilde{\lambda}}) - 1| \right\} \leq 1.$$

Based on the definition of effective dimension (Definition 5) and condition of bounded statistical leverage (Assumption 6), we can bound the smoothness as

$$\begin{aligned}
\max_{i \in [n]} x_i^\top \tilde{H}_\lambda^{-1} x_i &\leq 2 \max_{i \in [n]} x_i^\top H_\lambda^{-1} x_i \leq 2\rho_\lambda^2 \left(\frac{1}{n} \sum_{i=1}^n x_i^\top H_\lambda^{-1} x_i \right) \\
&= 2\rho_\lambda^2 \left(\frac{1}{n} \sum_{i=1}^n \text{tr}(x_i^\top H_\lambda^{-1} x_i) \right) \\
&\stackrel{\textcircled{1}}{=} 3\rho_\lambda^2 \left(\frac{1}{n} \sum_{i=1}^n \text{tr}(x_i x_i^\top H_\lambda^{-1}) \right) \\
&= 2\rho_\lambda^2 \left(\text{tr} \left(\frac{1}{n} \sum_i x_i x_i^\top H_\lambda^{-1} \right) \right) \\
&= 2\rho_\lambda^2 \left(\text{tr} \left(\left(\frac{1}{n} \sum_i x_i x_i^\top \right) H_\lambda^{-1} \right) \right) \\
&= 2\rho_\lambda^2 \sum_{j=1}^d \frac{\lambda_j}{\lambda_j + \tilde{\lambda}} = 3\rho_\lambda^2 d_{\tilde{\lambda}},
\end{aligned}$$

which finished the proof, where in step $\textcircled{1}$ we have used the fact that $\text{tr}(ABC) = \text{tr}(CAB)$ for any A, B, C . \square

A.6 Proof of Theorem 10

Proof. Since we choose $\tilde{\lambda}$ at the scale of $\tilde{\lambda} = \max \{ \lambda, \frac{L}{b} \}$, then we can apply Lemma 8 and Lemma 9 to verify that the new condition number after “preconditioning” will be

$$\max \left\{ 3\rho_\lambda^2 d_{\tilde{\lambda}}, \frac{3L\rho_\lambda^2 d_{\tilde{\lambda}}}{\lambda b} \right\},$$

applying Corollary 13 we know as long as the inexactness condition (40) is satisfied, we requires $\mathcal{O}(\log(\frac{1}{\epsilon}))$ full gradient evaluations and

$$\max \{ C_{\text{sb}}, C_{\text{lb}} \} \cdot \frac{10}{9} \log(36) \log \left(\frac{1}{\epsilon} \right) \quad (25)$$

total calls of approximate minimization of (4) to ensure $\mathbb{E}f(\tilde{w}_s) - f(w^*) \leq \epsilon$, where the factors $C_{\text{sb}}, C_{\text{lb}}$ are

$$\begin{aligned}
C_{\text{sb}} &= \max \left\{ \frac{240}{b}, \frac{240L}{\lambda b^2} \right\} \cdot \rho_\lambda^2 d_{\tilde{\lambda}}, \\
C_{\text{lb}} &= \max \left\{ 2\sqrt{6}, 2\sqrt{6} \sqrt{\frac{L}{\lambda b}} \right\} \cdot \sqrt{\rho_\lambda^2 d_{\tilde{\lambda}}},
\end{aligned}$$

which represent the cases of small and large minibatch sizes, respectively. Since the condition number of (4) is never larger than $L \cdot \frac{b}{L} = b$, we know from Lemma 23, when applying SVRG to solve (4), to reach some point of which objective the suboptimality of (4) satisfies

$$w_t - \min_w \tilde{f}_t(w) \leq \frac{1}{10^5} \cdot \left(\frac{\lambda}{L} \right)^7 \epsilon, \quad (26)$$

the following number of gradient calls sufficient:

$$C \cdot b \cdot \log\left(\frac{L}{\lambda}\right) \cdot \log\left(\frac{1}{\epsilon}\right), \quad (27)$$

for some universal constant C . On the other hand, for every subproblem (4), if (26) is satisfied, applying Lemma 4 we know the gradient error in IMBA-SVRG (Algorithm 3) in the preconditioned space is upper bounded by

$$\|\xi_t\|^2 \leq \left(\frac{\lambda}{L}\right)^6 \cdot \frac{2\epsilon^2}{10^5},$$

thus condition (40) in Theorem 12 is satisfied. Combining (25) and (27) we know the total number of gradient calls:

$$n \cdot \log\left(\frac{1}{\epsilon}\right) + \frac{10C}{9} \log(36) \cdot \max\{C_{\text{sb}}, C_{\text{lb}}\} \cdot b \cdot \log\left(\frac{L}{\lambda}\right) \cdot \log^2\left(\frac{1}{\epsilon}\right) \quad (28)$$

is sufficient to obtain a solution such that $\mathbb{E}f(\tilde{w}_s) - f(w^*) \leq \epsilon$ is satisfied. Next we choose b such that the total iteration complexity of above expression is minimized. Start from here we will ignore the constants before these factors for simplicity, we know the term $\max\{C_{\text{sb}}, C_{\text{lb}}\} \cdot b$ is of order

$$\mathcal{O}\left(\max\left\{\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}}, \frac{\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}} L}{\lambda b}, \sqrt{\frac{\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}} L b}{\lambda}}\right\}\right),$$

when $n \gtrsim \rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}} \left(\frac{L}{\lambda}\right)^{1/3}$, we can choose $b \asymp \rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}} \left(\frac{L}{\lambda}\right)^{1/3}$, then $\max\{C_{\text{sb}}, C_{\text{lb}}\} \cdot b$ is of order

$$\mathcal{O}\left(\max\left\{\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}}, \left(\frac{L}{\lambda}\right)^{2/3}, \rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}} \left(\frac{L}{\lambda}\right)^{2/3}\right\}\right) = \mathcal{O}\left(\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}} \left(\frac{L}{\lambda}\right)^{2/3}\right), \quad (29)$$

when $n \lesssim \rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}} \left(\frac{L}{\lambda}\right)^{1/3}$, we can choose $b \asymp n$, in which case $\max\{C_{\text{sb}}, C_{\text{lb}}\} \cdot b$ can be upper bounded by

$$\mathcal{O}\left(\max\left\{\frac{\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}} L}{\lambda n}, \sqrt{\frac{\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}} L n}{\lambda}}\right\}\right) = \mathcal{O}\left(\frac{\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}} L}{\lambda n}\right), \quad (30)$$

Combining (28), (29) and (30), we know the total number of individual function gradient calls to reach ϵ -suboptimality is

$$\mathcal{O}\left(\tilde{\kappa} \cdot \log\left(\frac{L}{\lambda}\right) \cdot \log^2\left(\frac{1}{\epsilon}\right) + n \cdot \log\left(\frac{1}{\epsilon}\right)\right),$$

where

$$\tilde{\kappa} = \max\left\{\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}} \left(\frac{L}{\lambda}\right)^{2/3}, \frac{\rho_{\tilde{\lambda}}^2 d_{\tilde{\lambda}} L}{\lambda n}\right\},$$

which finishes the proof. \square

A.7 Proof of Theorem 11

Proof. Applying the theory of catalyst acceleration (Lemma 24) we know only $\mathcal{O}\left(\left(\sqrt{\frac{\lambda+\gamma}{\lambda}}\right) \log\left(\frac{1}{\epsilon}\right)\right)$ calls of MP-SVRP is sufficient to reach ϵ -objective suboptimality, as long as each iterate w_r satisfies

$$f(w_r) + \frac{\gamma}{2} \|w_r - z_{r-1}\|^2 \leq \min_w f(w) + \frac{\gamma}{2} \|w - z_{r-1}\|^2 + \frac{\lambda\epsilon}{3600(\lambda+\gamma)} \left(1 - \frac{9}{10} \sqrt{\frac{\lambda}{\lambda+\gamma}}\right).$$

Moreover, according to Theorem 10, the iteration complexity of solving a $\lambda + \gamma$ strongly convex problem (17) is

$$\tilde{\mathcal{O}}\left(n + \max\left\{\rho_{\lambda}^2 d_{\tilde{\lambda}} \left(\frac{L}{(\lambda+\gamma)}\right)^{2/3}, \frac{\rho_{\lambda}^2 d_{\tilde{\lambda}} L}{(\lambda+\gamma)n}\right\}\right),$$

combining these two results we know the total iteration complexity of Algorithm 2 can be upper bounded by

$$\tilde{\mathcal{O}}\left(\left(\sqrt{\frac{\lambda+\gamma}{\lambda}}\right) \left(n + \max\left\{\left(\frac{\rho_{\lambda}^2 d_{\tilde{\lambda}} L}{(\lambda+\gamma)}\right)^{2/3}, \frac{\rho_{\lambda}^2 d_{\tilde{\lambda}} L}{(\lambda+\gamma)n}\right\}\right)\right),$$

When $\frac{\rho_{\lambda}^2 d_{\tilde{\lambda}} L}{\lambda} \geq n^{3/2}$, if we choose

$$\gamma = \frac{\rho_{\lambda}^2 d_{\tilde{\lambda}} L}{n^{3/2}} - \lambda,$$

we obtain the iteration complexity can be upper bounded by

$$\tilde{\mathcal{O}}\left(\left(\sqrt{\frac{\rho_{\lambda}^2 d_{\tilde{\lambda}} L}{n^{3/2}\lambda}}\right) \cdot n\right) = \tilde{\mathcal{O}}\left(\sqrt{\rho_{\lambda}^2 d_{\tilde{\lambda}}} \cdot n^{1/4} \left(\frac{L}{\lambda}\right)^{1/2}\right),$$

which concludes the proof. \square

B Proofs in Section 4

B.1 Proof of Lemma 15

Proof. Since

$$\begin{aligned} \mathbb{E}[f(w_t)] &\leq \min_w \{\mathbb{E}[V_t(w)]\} + \mathbb{E}[\delta_t] \\ &\leq \mathbb{E}[V_t(w^*)] + \mathbb{E}[\delta_t] \\ &\leq (1 - \theta_t)f(w^*) + \theta_t V_0(w^*) + \mathbb{E}[\delta_t], \end{aligned}$$

subtracting both sides by $f(w^*)$ finishes the proof. \square

B.2 Proof of Lemma 16

Proof. We proceed the proof by induction. When $t = 0$, we have

$$V_0(w) = (1 - \theta_0)f(w) + \theta_0 V_0(w) = V_0(w),$$

suppose

$$\mathbb{E}[V_{t-1}(w)] \leq (1 - \theta_{t-1})f(w) + \theta_{t-1}V_0(w)$$

is true for some $t \geq 1$, then for $\mathbb{E}[V_t(w)]$, since $\mathbb{E}[v_t] = \nabla f(y_{t-1})$, we have

$$\begin{aligned} \mathbb{E}[V_t(w)] &= (1 - \alpha_{t-1})\mathbb{E}[V_{t-1}(w)] \\ &\quad + \alpha_{t-1}\mathbb{E}\left[\left(\frac{1}{b}\sum_{i \in B_t} f_i(y_{t-1}) + \left\langle v_t - \frac{\xi_t}{\eta}, w - y_{t-1} \right\rangle + \frac{\lambda}{4}\|w - y_{t-1}\|^2 - \frac{\|\xi_t\|^2}{\lambda\eta^2}\right)\right] \\ &= (1 - \alpha_{t-1})\mathbb{E}[V_{t-1}(w)] \\ &\quad + \alpha_{t-1}\left(f(y_{t-1}) + \langle \nabla f(y_{t-1}), w - y_{t-1} \rangle + \frac{\lambda}{4}\|w - y_{t-1}\|^2 - \left\langle \frac{\xi_t}{\eta}, w - y_{t-1} \right\rangle - \frac{\|\xi_t\|^2}{\lambda\eta^2}\right) \\ &\stackrel{\textcircled{1}}{\leq} (1 - \alpha_{t-1})\mathbb{E}[V_{t-1}(w)] + \alpha_{t-1}\left(f(y_{t-1}) + \langle \nabla f(y_{t-1}), w - y_{t-1} \rangle + \frac{\lambda}{2}\|w - y_{t-1}\|^2\right) \\ &\stackrel{\textcircled{2}}{\leq} (1 - \alpha_{t-1})(1 - \theta_{t-1})f(w) + (1 - \alpha_{t-1})\theta_{t-1}V_0(w) + \alpha_{t-1}f(w) \\ &= (1 - (1 - \alpha_{t-1})\theta_{t-1})f(w) + (1 - \alpha_{t-1})\theta_{t-1}V_0(w) \\ &= (1 - \theta_t)f(w) + \theta_tV_0(w), \end{aligned}$$

which concludes the proof, where at step ① we used the inequality $-\langle v_1, v_2 \rangle - \|v_2\|^2 \leq \frac{\|v_1\|^2}{4}$, at step ② we used the inductive hypothesis and the fact that $f(w)$ is λ -strongly convex. \square

B.3 Proof of Lemma 17

Proof. We proceed the proof by induction, when $t = 0$ this is true by construction. Suppose for some $t \geq 1$ the following holds:

$$V_{t-1}(w) = V_{t-1}^* + \frac{\lambda}{4}\|w - z_{t-1}\|^2,$$

then at time t , we have

$$\begin{aligned} V_t(w) &= (1 - \alpha_{t-1})V_{t-1}(w) \\ &\quad + \alpha_{t-1}\left(\frac{1}{b}\sum_{i \in B_t} f_i(y_{t-1}) + \left\langle v_t - \frac{\xi_t}{\eta}, w - y_{t-1} \right\rangle + \frac{\lambda}{4}\|w - y_{t-1}\|^2 - \frac{\|\xi_t\|^2}{\lambda\eta^2}\right) \\ &= (1 - \alpha_{t-1})V_{t-1}^* + \frac{(1 - \alpha_{t-1})\lambda}{4}\|w - z_{t-1}\|^2 + \frac{\alpha_{t-1}\lambda}{4}\|w - y_{t-1}\|^2 + \alpha_{t-1}\left\langle v_t - \frac{\xi_t}{\eta}, w \right\rangle \\ &\quad + \frac{\alpha_{t-1}}{b}\sum_{i \in B_t} f_i(y_{t-1}) - \alpha_{t-1}\left\langle v_t - \frac{\xi_t}{\eta}, y_{t-1} \right\rangle - \frac{\alpha_{t-1}\|\xi_t\|^2}{\lambda\eta^2}, \end{aligned}$$

by first order optimality condition, it is clear that the minimizer of $V_t(w)$: z_t satisfies the following:

$$(1 - \alpha_{t-1})\lambda(z_t - z_{t-1}) + \alpha_{t-1}\lambda(z_t - y_{t-1}) + 2\alpha_{t-1}\left(v_t - \frac{\xi_t}{\eta}\right) = 0,$$

from above we obtain the recursive form of z_t as (23) defines. Plug in (23) to (21) we get

$$\begin{aligned}
V_t^* &= V_t(z_t) = (1 - \alpha_{t-1})V_{t-1}^* + \frac{(1 - \alpha_{t-1})\lambda}{4} \|z_t - z_{t-1}\|^2 + \frac{\alpha_{t-1}\lambda}{4} \|z_t - y_{t-1}\|^2 + \alpha_{t-1} \left\langle v_t - \frac{\xi_t}{\eta}, z_t \right\rangle \\
&\quad + \frac{\alpha_{t-1}}{b} \sum_{i \in B_t} f_i(y_{t-1}) - \alpha_{t-1} \left\langle v_t - \frac{\xi_t}{\eta}, y_{t-1} \right\rangle - \frac{\alpha_{t-1} \|\xi_t\|^2}{\lambda \eta^2} \\
&= (1 - \alpha_{t-1})V_{t-1}^* + \frac{(1 - \alpha_{t-1})\lambda}{4} \left\| \alpha_{t-1}(y_{t-1} - z_{t-1}) - \frac{2\alpha_{t-1}}{\lambda} \left(v_t - \frac{\xi_t}{\eta} \right) \right\|^2 + \frac{\alpha_{t-1}}{b} \sum_{i \in B_t} f_i(y_{t-1}) \\
&\quad + \frac{\alpha_{t-1}\lambda}{4} \left\| (1 - \alpha_{t-1})(z_{t-1} - y_{t-1}) - \frac{2\alpha_{t-1}}{\lambda} \left(v_t - \frac{\xi_t}{\eta} \right) \right\|^2 - \frac{\alpha_{t-1} \|\xi_t\|^2}{\lambda \eta^2} \\
&\quad + \alpha_{t-1} \left\langle v_t - \frac{\xi_t}{\eta}, (1 - \alpha_{t-1})(z_{t-1} - y_{t-1}) - \frac{2\alpha_{t-1}}{\lambda} \left(v_t - \frac{\xi_t}{\eta} \right) \right\rangle \\
&= (1 - \alpha_{t-1})V_{t-1}^* + \frac{\alpha_{t-1}(1 - \alpha_{t-1})\lambda}{4} \|z_{t-1} - y_{t-1}\|^2 + \alpha_{t-1}(1 - \alpha_{t-1}) \left\langle v_t - \frac{\xi_t}{\eta}, z_{t-1} - y_{t-1} \right\rangle \\
&\quad - \frac{\alpha_{t-1}^2}{\lambda} \|v_t\|^2 + \frac{\alpha_{t-1}}{b} \sum_{i \in B_t} f_i(y_{t-1}) - \frac{(\alpha_{t-1} + \alpha_{t-1}^2) \|\xi_t\|^2}{\lambda \eta^2} + \frac{2\alpha_{t-1}^2}{\lambda \eta} \langle \xi_t, v_t \rangle,
\end{aligned}$$

which verified (24). \square

B.4 Proof of Lemma 18

Proof. We prove by induction, when $t = 0$ it is obviously true, suppose it is true for $t - 1, \forall t \geq 1$, i.e.

$$z_{t-1} - y_{t-1} = \frac{1}{\alpha}(y_{t-1} - w_{t-1}),$$

for iteration t , based on (23) we have

$$\begin{aligned}
z_t - y_t &= (1 - \alpha)z_{t-1} + \alpha y_{t-1} - \frac{2\alpha}{\lambda} \left(v_t - \frac{\xi_t}{\eta} \right) - y_t \\
&= (1 - \alpha)(z_{t-1} - y_{t-1}) + y_{t-1} - \frac{2\alpha}{\lambda} \left(v_t - \frac{\xi_t}{\eta} \right) - y_t \\
&\stackrel{\textcircled{1}}{=} \frac{1 - \alpha}{\alpha}(y_{t-1} - w_{t-1}) + y_{t-1} - \frac{2\alpha}{\lambda} \left(v_t - \frac{\xi_t}{\eta} \right) - y_t \\
&= \frac{1}{\alpha} \left(y_{t-1} - \frac{2\alpha^2}{\lambda} \left(v_t - \frac{\xi_t}{\eta} \right) \right) - \frac{1 - \alpha}{\alpha} w_{t-1} - y_t \\
&\stackrel{\textcircled{2}}{=} \frac{1}{\alpha} w_t - \frac{1 - \alpha}{\alpha} w_{t-1} - y_t \\
&\stackrel{\textcircled{3}}{=} \frac{1}{\alpha} \left(w_t + \frac{1 - \alpha}{1 + \alpha} (w_t - w_{t-1}) - w_t \right) \\
&\stackrel{\textcircled{4}}{=} \frac{1}{\alpha} (y_t - w_t),
\end{aligned}$$

which concludes the proof, where step ① used the inductive hypothesis, step ② used the update rule of w_t in Algorithm 3, step ③ and ④ used the update rule of y_t in Algorithm 3. \square

B.5 Proof of Lemma 19

Proof. We prove by induction, when $t = 0$, it is true that

$$f(w_0) \leq V_0^* = f(z_0) = f(w_0),$$

suppose

$$\mathbb{E}[f(w_{t-1})] \leq \mathbb{E}[V_{t-1}^*] + \mathbb{E}[\delta_{t-1}], \quad (31)$$

for some $t \geq 1$, then based on smoothness, we know

$$f(w_t) \leq f(y_{t-1}) + \langle \nabla f(y_{t-1}), w_t - y_{t-1} \rangle + \frac{L}{2} \|w_t - y_{t-1}\|^2,$$

thus

$$\begin{aligned} V_t^* = & (1 - \alpha_{t-1})V_{t-1}^* + \frac{\alpha_{t-1}(1 - \alpha_{t-1})\lambda}{4} \|z_{t-1} - y_{t-1}\|^2 + \alpha_{t-1}(1 - \alpha_{t-1}) \left\langle v_t - \frac{\xi_t}{\eta}, z_{t-1} - y_{t-1} \right\rangle \\ & - \frac{\alpha_{t-1}^2}{\lambda} \|v_t\|^2 + \frac{\alpha_{t-1}}{b} \sum_{i \in B_t} f_i(y_{t-1}) - \frac{(\alpha_{t-1} + \alpha_{t-1}^2) \|\xi_t\|^2}{\lambda \eta^2} + \frac{2\alpha_{t-1}^2}{\lambda \eta} \langle \xi_t, v_t \rangle. \end{aligned}$$

$$\begin{aligned} \mathbb{E}[f(w_t) - V_t^*] \leq & \mathbb{E} \left[f(y_{t-1}) + \langle \nabla f(y_{t-1}), w_t - y_{t-1} \rangle + \frac{L}{2} \|w_t - y_{t-1}\|^2 \right] \\ & - \mathbb{E} \left[(1 - \alpha)V_{t-1}^* - \frac{\alpha(1 - \alpha)\lambda}{4} \|z_{t-1} - y_{t-1}\|^2 - \alpha(1 - \alpha) \left\langle v_t - \frac{\xi_t}{\eta}, z_{t-1} - y_{t-1} \right\rangle \right] \\ & + \mathbb{E} \left[\frac{\alpha^2}{\lambda} \|v_t\|^2 - \frac{\alpha}{b} \sum_{i \in B_t} f_i(y_{t-1}) + \frac{(\alpha + \alpha^2) \|\xi_t\|^2}{\lambda \eta^2} - \frac{2\alpha^2}{\lambda \eta} \langle \xi_t, v_t \rangle \right] \\ \stackrel{\textcircled{1}}{=} & \mathbb{E} \left[(1 - \alpha)(f(y_{t-1}) - V_{t-1}^* + \langle v_t, w_{t-1} - y_{t-1} \rangle) - \langle \nabla f(y_{t-1}), \eta v_t - \xi_t \rangle + \frac{\eta}{2} \|v_t\|^2 \right] \\ & + \mathbb{E} \left[\frac{L}{2} \|\eta v_t - \xi_t\|^2 - \frac{(1 - \alpha)\lambda}{4\alpha} \|y_{t-1} - w_{t-1}\|^2 + \frac{(\alpha + \alpha^2) \|\xi_t\|^2}{\lambda \eta^2} \right] \\ & + \mathbb{E} \left[-\frac{2\alpha^2}{\lambda \eta} \langle \xi_t, v_t \rangle - (1 - \alpha) \left\langle \frac{\xi_t}{\eta}, w_{t-1} - y_{t-1} \right\rangle \right] \\ \stackrel{\textcircled{2}}{\leq} & \mathbb{E} \left[(1 - \alpha)(f(y_{t-1}) - V_{t-1}^* + \langle v_t, w_{t-1} - y_{t-1} \rangle) - \langle \nabla f(y_{t-1}), \eta v_t - \xi_t \rangle + \frac{9\eta}{16} \|v_t\|^2 \right] \\ & + \mathbb{E} \left[\frac{L}{2} \|\eta v_t - \xi_t\|^2 + \left(\frac{2\alpha}{\lambda \eta^2} + \frac{16\alpha^4}{\lambda^2 \eta^3} + \frac{(1 - \alpha)}{2\lambda \eta^2} \right) \|\xi_t\|^2 \right] \\ & + \mathbb{E} \left[\left(\frac{(1 - \alpha)\lambda}{2} - \frac{(1 - \alpha)\lambda}{4\alpha} \right) \|w_{t-1} - y_{t-1}\|^2 \right] \quad (32) \end{aligned}$$

where at step $\textcircled{1}$ we used the fact of updating rule

$$w_t = y_{t-1} - \eta v_t + \xi_t,$$

as well as Lemma 18, and $\mathbb{E} \left[\frac{1}{b} \sum_{i \in B_t} f_i(y_{t-1}) \right] = f(y_{t-1})$; and at step ② we used the following two inequalities:

$$\begin{aligned} \left| (1-\alpha) \left\langle \frac{\xi_t}{\eta}, w_{t-1} - y_{t-1} \right\rangle \right| &\leq \frac{(1-\alpha)\lambda}{2} \|y_{t-1} - w_{t-1}\|^2 + \frac{(1-\alpha)}{2\lambda\eta^2} \|\xi_t\|^2, \\ \left| \frac{2\alpha^2}{\lambda\eta} \langle \xi_t, v_t \rangle \right| &\leq \frac{\eta}{16} \|v_t\|^2 + \frac{16\alpha^4}{\lambda^2\eta^3} \|\xi_t\|^2 \end{aligned}$$

Since

$$\begin{aligned} \mathbb{E} [f(y_{t-1}) - V_{t-1}^* + \langle v_t, w_{t-1} - y_{t-1} \rangle] &= \mathbb{E} [f(y_{t-1}) - V_{t-1}^* + \langle \nabla f(y_{t-1}), w_{t-1} - y_{t-1} \rangle] \\ &\stackrel{\textcircled{1}}{\leq} \mathbb{E} \left[f(w_{t-1}) - \frac{\lambda}{2} \|w_{t-1} - y_{t-1}\|^2 - V_{t-1}^* \right] \\ &\stackrel{\textcircled{2}}{\leq} \mathbb{E} \left[\delta_{t-1} - \frac{\lambda}{2} \|w_{t-1} - y_{t-1}\|^2 \right], \end{aligned} \quad (33)$$

where at step ① we used the λ -strong convexity of $f(w)$, at step ② we used the inductive hypothesis (31). Combining (32) and (33) we obtain

$$\begin{aligned} \mathbb{E} [f(w_t) - V_t^*] &\leq (1-\alpha)\delta_{t-1} - \mathbb{E} \left[\frac{(1-\alpha)\lambda}{4\alpha} \|y_{t-1} - w_{t-1}\|^2 \right] \\ &\quad + \mathbb{E} \left[\frac{L}{2} \|\eta v_t - \xi_t\|^2 + \frac{9\eta}{16} \|v_t\|^2 - \langle \nabla f(y_{t-1}), \eta v_t - \xi_t \rangle + \left(\frac{2\alpha}{\lambda\eta^2} + \frac{16\alpha^4}{\lambda^2\eta^3} + \frac{(1-\alpha)}{2\lambda\eta^2} \right) \|\xi_t\|^2 \right]. \end{aligned} \quad (34)$$

Next we bound the third term on the right hand side, since

$$\begin{aligned} \frac{L}{2} \|\eta v_t - \xi_t\|^2 + \frac{9\eta}{16} \|v_t\|^2 - \langle \nabla f(y_{t-1}), \eta v_t + \xi_t \rangle &\stackrel{\textcircled{1}}{\leq} L\eta^2 \|v_t\|^2 + L \|\xi_t\|^2 + \frac{9\eta}{16} \|v_t\|^2 - \langle \nabla f(y_{t-1}), \eta v_t - \xi_t \rangle \\ &= \frac{\eta}{2} \|v_t - \nabla f(y_{t-1})\|^2 - \frac{\eta}{2} \|\nabla f(y_{t-1})\|^2 \\ &\quad + \left(L\eta^2 + \frac{\eta}{16} \right) \|v_t\|^2 + \langle \nabla f(y_{t-1}), \xi_t \rangle + L \|\xi_t\|^2 \\ &\stackrel{\textcircled{2}}{\leq} \frac{\eta}{2} \|v_t - \nabla f(y_{t-1})\|^2 - \frac{\eta}{2} \|\nabla f(y_{t-1})\|^2 \\ &\quad + \left(2L\eta^2 + \frac{\eta}{8} \right) (\|v_t - \nabla f(y_{t-1})\|^2 + \|\nabla f(y_{t-1})\|^2) \\ &\quad + \frac{\eta}{8} \|\nabla f(y_{t-1})\|^2 + \frac{2}{\eta} \|\xi_t\|^2 + L \|\xi_t\|^2, \end{aligned}$$

where at step ① we used $\|v_1 + v_2\|^2 \leq 2(\|v_1\|^2 + \|v_2\|^2)$, and at step ② we used it again and also the inequality $2|\langle v_1, v_2 \rangle| \leq \|v_1\|^2 + \|v_2\|^2$. Thus we know when $\eta \leq \frac{1}{8L}$, we have

$$\begin{aligned} \frac{L}{2} \|\eta v_t - \xi_t\|^2 + \frac{9\eta}{16} \|v_t\|^2 - \langle \nabla f(y_{t-1}), \eta v_t + \xi_t \rangle &\leq \left(\frac{\eta}{2} + 2L\eta^2 + \frac{\eta}{8} \right) \|v_t - \nabla f(y_{t-1})\|^2 + \left(L + \frac{2}{\eta} \right) \|\xi_t\|^2 \\ &\quad - \left(\frac{\eta}{2} - \frac{\eta}{8} - \frac{\eta}{8} - 2L\eta^2 \right) \|\nabla f(y_{t-1})\|^2 \\ &\leq \eta \|v_t - \nabla f(y_{t-1})\|^2 + \frac{3}{\eta} \|\xi_t\|^2. \end{aligned} \quad (35)$$

Combining (34) and (35) we obtain

$$\begin{aligned}\mathbb{E}[f(w_t) - V_t^*] &\leq (1 - \alpha)\delta_{t-1} + \mathbb{E}\left[\eta\|v_t - \nabla f(y_{t-1})\|^2 - \frac{(1 - \alpha)\lambda}{4\alpha}\|y_{t-1} - w_{t-1}\|^2\right] \\ &\quad + \mathbb{E}\left[\left(\frac{2\alpha}{\lambda\eta^2} + \frac{16\alpha^4}{\lambda^2\eta^3} + \frac{(1 - \alpha)}{2\lambda\eta^2} + \frac{3}{\eta}\right)\|\xi_t\|^2\right],\end{aligned}\tag{36}$$

also since

$$\alpha = \sqrt{\frac{\lambda\eta}{2}} \leq \frac{1}{4}\sqrt{\frac{\lambda}{L}} \leq \frac{1}{4},$$

we know

$$\frac{2\alpha}{\lambda\eta^2} + \frac{16\alpha^4}{\lambda^2\eta^3} + \frac{(1 - \alpha)}{2\lambda\eta^2} + \frac{3}{\eta} \leq \frac{1}{16\lambda^2\eta^3} + \frac{1}{16\lambda^2\eta^3} + \frac{1}{16\lambda^2\eta^3} + \frac{3}{64\lambda^2\eta^3} \leq \frac{1}{4\lambda^2\eta^3},$$

combining the inequality above and (36), then substituting $\alpha = \sqrt{\frac{\lambda\eta}{2}}$ finishes the proof. \square

B.6 Proof of Lemma 20

Proof. The proof follows the strategy in (Johnson and Zhang, 2013; Nitanda, 2014). First, based on the minibatch sampling (here for simplicity we only consider sampling with replacement, for sampling without replacement, the bound below can be tightened by a factor of $\frac{n-b}{n-1}$, see, e.g. Section 2.8 of (Lohr, 2009)) we know

$$\begin{aligned}\mathbb{E}\left[\|v_t - \nabla f(y_{t-1})\|^2\right] &= \frac{1}{b^2}\mathbb{E}\left[\sum_{j \in B_t} \|\nabla f_j(y_{t-1}) - \nabla f(y_{t-1}) - (\nabla f_j(\tilde{w}_{s-1}) - \nabla f(\tilde{w}_{s-1}))\|^2\right] \\ &\stackrel{\textcircled{1}}{\leq} \frac{1}{b^2}\mathbb{E}\left[\sum_{j \in B_t} \|\nabla f_j(y_{t-1}) - \nabla f_j(\tilde{w}_{s-1})\|^2\right] \\ &\stackrel{\textcircled{2}}{\leq} \frac{2}{b^2}\mathbb{E}\left[\sum_{j \in B_t} (\|\nabla f_j(y_{t-1}) - \nabla f_j(w_{t-1})\|^2 + \|\nabla f_j(\tilde{w}_{s-1}) - \nabla f_j(w_{t-1})\|^2)\right] \\ &\stackrel{\textcircled{3}}{\leq} \frac{4}{b^2}\mathbb{E}\left[\sum_{j \in B_t} (\|\nabla f_j(\tilde{w}_{s-1}) - \nabla f_j(w^*)\|^2 + \|\nabla f_j(w^*) - \nabla f_j(w_{t-1})\|^2)\right] \\ &\quad + \frac{2L^2}{b}\|y_{t-1} - w_{t-1}\|^2 \\ &\stackrel{\textcircled{4}}{\leq} \frac{8L}{b}(f(w_{t-1}) - f(w^*) + f(\tilde{w}_{s-1}) - f(w^*)) + \frac{2L^2}{b}\|y_{t-1} - w_{t-1}\|^2,\end{aligned}$$

where at step ① we used $\mathbb{E}\|v - \mathbb{E}v\|^2 \leq \mathbb{E}\|v\|^2$, at step ② we used $\|v_1 + v_2\|^2 \leq 2(\|v_1\|^2 + \|v_2\|^2)$, at step ③ we used it again along with the L -smoothness of $f_j(w)$, at step ④ we used standard results in SVRG analysis, e.g. Lemma 1 of (Xiao and Zhang, 2014). Substituting above into the term

$$\mathbb{E}\left[\eta\|v_t - \nabla f(y_{t-1})\|^2 - \frac{(1 - \sqrt{\lambda\eta/2})\lambda}{\sqrt{2\lambda\eta}}\|y_{t-1} - w_{t-1}\|^2\right] \text{ concludes the proof. } \square$$

B.7 Proof of Lemma 21

Proof. Combining Lemma 19 and 20, we have

$$\begin{aligned}
\delta_t &= \left(1 - \sqrt{\frac{\lambda\eta}{2}}\right) \delta_{t-1} + \left(\eta \|v_t - \nabla f(y_{t-1})\|^2 - \frac{(1 - \sqrt{\lambda\eta/2})\lambda}{\sqrt{2\lambda\eta}} \|y_{t-1} - w_{t-1}\|^2 + \frac{1}{4\lambda^2\eta^3} \|\xi_t\|^2\right) \\
&\leq \left(1 - \sqrt{\frac{\lambda\eta}{2}}\right) \delta_{t-1} + \left(\frac{1}{4\lambda^2\eta^3} \|\xi_t\|^2 + \frac{8\eta L}{b} (f(w_{t-1}) + f(\tilde{w}_{s-1}) - 2f(w^*))\right) \\
&\quad + \left(\frac{2\eta L^2}{b} - \frac{(1 - \sqrt{\lambda\eta/2})\lambda}{\sqrt{2\lambda\eta}}\right) \|y_{t-1} - w_{t-1}\|^2.
\end{aligned} \tag{37}$$

First we verify the factor before $\|y_{t-1} - w_{t-1}\|^2$ is non positive, since $1 - \eta\lambda \geq 1 - \frac{\lambda}{8L} \geq \frac{1}{2}$, then

$$\frac{2\eta L^2}{b} \cdot \frac{\sqrt{2\lambda\eta}}{(1 - \sqrt{\eta\lambda/2})\lambda} \leq \frac{4\eta^{3/2}L^2}{b\sqrt{\lambda}} \leq \min \left\{ \frac{b^2\lambda}{128000L}, \frac{L^{1/2}}{4b\sqrt{\lambda}} \right\} \leq 1,$$

where the last inequality is true because

$$\frac{b^2\lambda}{128000L} \left(\frac{L^{1/2}}{4b\sqrt{\lambda}} \right)^2 \leq 1,$$

thus we must have $\min \left\{ \frac{b^2\lambda}{128000L}, \frac{L^{1/2}}{4b\sqrt{\lambda}} \right\} \leq 1$ otherwise it leads to a contradiction. Thus $\frac{2\eta L^2}{b} - \frac{(1 - \sqrt{\eta\lambda/2})\lambda}{\sqrt{2\eta\lambda}} \leq 0$, combining with (37), we have

$$\delta_t \leq \left(1 - \sqrt{\frac{\lambda\eta}{2}}\right) \delta_{t-1} + \left(\frac{1}{4\lambda^2\eta^3} \|\xi_t\|^2 + \frac{8\eta L}{b} (f(w_{t-1}) + f(\tilde{w}_{s-1}) - 2f(w^*))\right),$$

applying above inequality recursively, we get

$$\delta_t \leq \sum_{k=1}^t \left(1 - \sqrt{\frac{\lambda\eta}{2}}\right)^{t-k} \left(\frac{1}{4\lambda^2\eta^3} \|\xi_k\|^2 + \frac{8\eta L}{b} (f(w_{k-1}) + f(\tilde{w}_{s-1}) - 2f(w^*))\right),$$

then applying Lemma 15 and Lemma 19 concludes the proof. \square

B.8 Proof of Theorem 12

Proof. We prove this theorem by induction, note that when $b \leq 20\sqrt{\frac{2L}{\lambda}}$, it is clear that

$$\frac{8\eta L}{b} \leq \frac{b^2\lambda}{6400L^2} \cdot \frac{8L}{b} = \frac{\lambda b}{800L} = \frac{\sqrt{\eta\lambda}}{10},$$

and when $b \geq 20\sqrt{\frac{2L}{\lambda}}$,

$$\frac{8\eta L}{b} = \frac{1}{b} \leq \frac{\sqrt{\lambda}}{20\sqrt{2L}} = \frac{\sqrt{\eta\lambda}}{10},$$

thus in both cases we have $\frac{8\eta L}{b} \leq \frac{\sqrt{\eta\lambda}}{10}$, thus

$$\sum_{k=1}^t \left(1 - \sqrt{\frac{\lambda\eta}{2}}\right)^{t-k} \left(\frac{8\eta L}{b}\right) \leq \frac{1}{10} \sum_{k=1}^t \left(1 - \sqrt{\frac{\lambda\eta}{2}}\right)^{t-k} \sqrt{\lambda\eta} \leq \frac{1}{10} \sum_{k=1}^{\infty} \left(1 - \sqrt{\frac{\lambda\eta}{2}}\right)^k \sqrt{\lambda\eta} \leq \frac{\sqrt{2}}{10}.$$

By Lemma 21 we know

$$\mathbb{E}f(w_t) - f(w^*) \leq \left(1 - \sqrt{\frac{\lambda\eta}{2}}\right)^t (V_0(w^*) - f(w^*)) + \frac{\sqrt{2}(f(\tilde{w}_{s-1}) - f(w^*))}{10} \quad (38)$$

$$+ \mathbb{E} \left[\sum_{k=1}^t \left(1 - \sqrt{\frac{\lambda\eta}{2}}\right)^{t-k} \left(\frac{1}{4\lambda^2\eta^3} \|\xi_k\|^2 + \frac{\sqrt{\lambda\eta}}{10} (f(w_{k-1}) - f(w^*)) \right) \right], \quad (39)$$

also when $\forall k \in [t]$,

$$\|\xi_k\|^2 \leq \frac{2\lambda^2\eta^3\sqrt{\lambda\eta}}{15} (f(\tilde{w}_{s-1}) - f(w^*)), \quad (40)$$

we know

$$\begin{aligned} \sum_{k=1}^t \left(1 - \sqrt{\frac{\lambda\eta}{2}}\right)^{t-k} \left(\frac{1}{4\lambda^2\eta^3} \|\xi_k\|^2 \right) &\leq \frac{1}{30} \sum_{k=1}^t \left(1 - \sqrt{\frac{\lambda\eta}{2}}\right)^{t-k} \sqrt{\lambda\eta} (f(\tilde{w}_{s-1}) - f(w^*)) \\ &\leq \frac{1}{30} \sum_{k=1}^{\infty} \left(1 - \sqrt{\frac{\lambda\eta}{2}}\right)^k \sqrt{\lambda\eta} (f(\tilde{w}_{s-1}) - f(w^*)) \\ &\leq \frac{(f(\tilde{w}_{s-1}) - f(w^*))}{20}. \end{aligned} \quad (41)$$

Combining (39) and (41), we have

$$\begin{aligned} \mathbb{E}f(w_t) - f(w^*) &\leq \left(1 - \sqrt{\frac{\lambda\eta}{2}}\right)^t (V_0(w^*) - f(w^*)) + \frac{(f(\tilde{w}_{s-1}) - f(w^*))}{5} \\ &\quad + \mathbb{E} \left[\sum_{k=1}^t \left(1 - \sqrt{\frac{\lambda\eta}{2}}\right)^{t-k} \left(\frac{\sqrt{\lambda\eta}}{10} (f(w_{k-1}) - f(w^*)) \right) \right], \end{aligned}$$

denote A_t be the right hand side on above inequality, i.e.

$$\begin{aligned}
A_t &= \left(1 - \sqrt{\frac{\lambda\eta}{2}}\right)^t (V_0(w^*) - f(w^*)) + \frac{(f(\tilde{w}_{s-1}) - f(w^*))}{5} \\
&\quad + \mathbb{E} \left[\sum_{k=1}^t \left(1 - \sqrt{\frac{\lambda\eta}{2}}\right)^{t-k} \left(\frac{\sqrt{\lambda\eta}}{10} (f(w_{k-1}) - f(w^*)) \right) \right] \\
&= \left(1 - \sqrt{\frac{\lambda\eta}{2}}\right) \left(\left(1 - \sqrt{\frac{\lambda\eta}{2}}\right)^{t-1} (V_0(w^*) - f(w^*)) + \frac{(f(\tilde{w}_{s-1}) - f(w^*))}{5} \right) \\
&\quad + \left(1 - \sqrt{\frac{\lambda\eta}{2}}\right) \mathbb{E} \left[\sum_{k=1}^{t-1} \left(1 - \sqrt{\frac{\lambda\eta}{2}}\right)^{t-1-k} \left(\frac{\sqrt{\lambda\eta}}{10} (f(w_{k-1}) - f(w^*)) \right) \right] \\
&\quad + \frac{\sqrt{\lambda\eta}}{5\sqrt{2}} (f(\tilde{w}_{s-1}) - f(w^*)) + \frac{\sqrt{\lambda\eta}}{10} (f(w_{t-1}) - f(w^*)) \\
&\stackrel{\textcircled{1}}{\leq} \left(1 - \sqrt{\frac{\lambda\eta}{2}}\right) A_{t-1} + \frac{\sqrt{\lambda\eta}}{5\sqrt{2}} A_0 + \frac{\sqrt{\lambda\eta}}{10} A_{t-1} \\
&= \left(1 - \frac{9\sqrt{\lambda\eta}}{10}\right) A_{t-1} + \frac{\sqrt{\lambda\eta}}{5\sqrt{2}} A_0,
\end{aligned}$$

where at step $\textcircled{1}$ we used the fact that $f(\tilde{w}_{s-1}) - f(w^*) \leq A_0$, and $f(w_{t-1}) - f(w^*) \leq A_{t-1}$, applying above inequality recursively we get

$$\begin{aligned}
A_t &\leq \left(1 - \frac{9\sqrt{\lambda\eta}}{10}\right)^t A_0 + \frac{\sqrt{\lambda\eta}}{5\sqrt{2}} \sum_{k=0}^{t-1} \left(1 - \frac{9\sqrt{\lambda\eta}}{10}\right)^k A_0 \\
&\leq \left(1 - \frac{9\sqrt{\lambda\eta}}{10}\right)^t A_0 + \frac{\sqrt{\lambda\eta}}{5\sqrt{2}} \sum_{k=0}^{\infty} \left(1 - \frac{9\sqrt{\lambda\eta}}{10}\right)^k A_0 \leq \left(\left(1 - \frac{9\sqrt{\lambda\eta}}{10}\right)^t + \frac{2}{9} \right) A_0,
\end{aligned}$$

thus we know when $t \geq \frac{10}{9\sqrt{\lambda\eta}} \log(36)$, we have

$$A_t \leq \left(\frac{1}{36} + \frac{2}{9} \right) A_0 = \frac{1}{4} A_0,$$

also since

$$A_0 = V_0(w^*) - f(w^*) = f(\tilde{w}_{s-1}) - f(w^*) + \frac{\lambda}{2} \|\tilde{w}_{s-1} - w^*\|^2 \stackrel{\textcircled{1}}{\leq} 2(f(\tilde{w}_{s-1}) - f(w^*)),$$

where at step $\textcircled{1}$ we used the λ -strong convexity of $f(w)$. Combine above two inequality we get when $t \geq \frac{10}{9\sqrt{\lambda\eta}} \log(36)$, the expected objective suboptimality is halved:

$$\mathbb{E}f(w_t) - f(w^*) \leq A_t \leq \frac{1}{4} A_0 \leq \frac{1}{2} (f(\tilde{w}_{s-1}) - f(w^*)),$$

which concludes the proof. \square

C Collections of tools in the analysis

Lemma 22. (*Matrix Bernstein, Theorem 1.4 of (Tropp, 2012) rephrased*) Let X_1, \dots, X_k be some independent, self-adjoint random matrices with dimension d , and assume each random matrix satisfies:

$$\mathbb{E}X_k = 0 \quad \text{and} \quad \|X_k\| \leq R \quad \text{almost surely.}$$

Then, for all $t \geq 0$,

$$P\left(\left\|\sum_k X_k\right\| \geq t\right) \leq d \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right)$$

where

$$\sigma^2 := \left\|\sum_k \mathbb{E}(X_k^2)\right\|.$$

Lemma 23. (*Iteration complexity of SVRG, Corollary 1 of (Xiao and Zhang, 2014) rephrased*) If we apply SVRG to any finite-sum optimization objective $f(w)$ where each individual function is λ -strongly convex and L -smooth, then there are universal constant C such that for any target accuracy ε , and success probability $1 - \delta$, SVRG is able to find a solution that satisfies ε objective suboptimality using

$$C \cdot \left(n + \frac{L}{\lambda}\right) \log\left(\frac{\varepsilon_0}{\delta\varepsilon}\right)$$

first order oracle calls of individual functions, where n is the total number of individual functions in $f(w)$, ε_0 is the initial objective suboptimality.

Lemma 24. (*Iteration complexity of catalyst acceleration, Theorem 3.1 of (Lin et al., 2015a) rephrased*) For any λ -strongly convex and L -smooth function $f(w)$, If the minimization step of (17) satisfies

$$f(w_r) - \frac{\gamma}{2} \|w_r - z_{r-1}\| - \min_w \left(f(w) + \frac{\gamma}{2} \|w - z_{r-1}\|\right) \leq \frac{2}{9} (f(w_0) - f(w^*)) \left(1 - \frac{9}{10} \sqrt{\frac{\lambda}{\lambda + \gamma}}\right)^r,$$

then if we initialize $\nu_0 = \sqrt{\frac{\lambda}{\lambda + \gamma}}$ and set ν_r such that $\nu_r^2 = (1 - \nu_r)\nu_{r-1}^2 + (\lambda\nu_r)/(\lambda + \gamma)$, then the sequences $\{w_r\}$ in Algorithm 2 satisfies

$$f(w_r) - f(w^*) \leq \frac{800(\lambda + \gamma)}{\lambda} \left(1 - \frac{9}{10} \sqrt{\frac{\lambda}{\lambda + \gamma}}\right)^{r+1} (f(w_0) - f(w^*)).$$

References

- N. Agarwal, B. Bullins, and E. Hazan. Second order stochastic optimization in linear time. *arXiv preprint arXiv:1602.03943*, 2016.
- Z. Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *arXiv preprint arXiv:1603.05953*, 2016.
- Y. Arjevani and O. Shamir. Oracle complexity of second-order methods for finite-sum problems. *arXiv preprint arXiv:1611.04982*, 2016.
- R. Bollapragada, R. Byrd, and J. Nocedal. Exact and inexact subsampled newton methods for optimization. *arXiv preprint arXiv:1609.08502*, 2016.
- L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *arXiv preprint arXiv:1606.04838*, 2016.
- R. H. Byrd, G. M. Chin, W. Neveitt, and J. Nocedal. On the use of stochastic hessian information in optimization methods for machine learning. *SIAM Journal on Optimization*, 21(3):977–995, 2011.
- R. H. Byrd, S. L. Hansen, J. Nocedal, and Y. Singer. A stochastic quasi-newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2):1008–1031, 2016.
- S. Chatterjee and A. S. Hadi. *Sensitivity analysis in linear regression*, volume 327. John Wiley & Sons, 2009.
- A. Cotter, O. Shamir, N. Srebro, and K. Sridharan. Better mini-batch algorithms via accelerated gradient methods. In *Advances in neural information processing systems*, pages 1647–1655, 2011.
- A. Defazio. A simple practical accelerated method for finite sums. In *Advances In Neural Information Processing Systems*, pages 676–684, 2016.
- 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*, pages 1646–1654, 2014.
- O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(Jan):165–202, 2012.
- O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1-2):37–75, 2014.
- M. A. Erdogdu and A. Montanari. Convergence rates of sub-sampled newton methods. In *Advances in Neural Information Processing Systems*, pages 3052–3060, 2015.
- R. Frostig, R. Ge, S. Kakade, and A. Sidford. Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 2540–2548, 2015.
- A. Gonen, F. Orabona, and S. Shalev-Shwartz. Solving ridge regression using sketched preconditioned svrg. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1397–1405, 2016.

- R. Gower, D. Goldfarb, and P. Richtárik. Stochastic block bfgs: squeezing more curvature out of data. In *International Conference on Machine Learning*, pages 1869–1878, 2016.
- D. Hsu, S. M. Kakade, and T. Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 3(14):569–600, 2014.
- R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.
- J. Konečný, 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(2):242–255, 2016.
- G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.
- M. Li, T. Zhang, Y. Chen, and A. J. Smola. Efficient mini-batch training for stochastic optimization. In *Proc. of the 20th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (SIGKDD 2014)*, pages 661–670, 2014.
- H. Lin, J. Mairal, and Z. Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, pages 3384–3392, 2015a.
- Q. Lin, X. Chen, and J. Peña. A sparsity preserving stochastic gradient methods for sparse regression. *Computational Optimization and Applications*, 58(2):455–482, 2014.
- Q. Lin, Z. Lu, and L. Xiao. An accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization. *SIAM Journal on Optimization*, 25(4):2244–2273, 2015b.
- D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.
- S. Lohr. *Sampling: design and analysis*. Nelson Education, 2009.
- M. W. Mahoney et al. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.
- P. Moritz, R. Nishihara, and M. Jordan. A linearly-convergent stochastic l-bfgs algorithm. In *Artificial Intelligence and Statistics*, pages 249–258, 2016.
- Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2004.
- A. Nitanda. Stochastic proximal gradient descent with acceleration techniques. In *Advances in Neural Information Processing Systems*, pages 1574–1582, 2014.
- M. Pilanci and M. J. Wainwright. Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1):205–245, 2017.
- Z. Qu, P. Richtárik, M. Takáč, and O. Fercoq. Sdna: Stochastic dual newton ascent for empirical risk minimization. *arXiv preprint arXiv:1502.02268*, 2015.
- F. Roosta-Khorasani and M. W. Mahoney. Sub-sampled newton methods i: globally convergent algorithms. *arXiv preprint arXiv:1601.04737*, 2016a.

- F. Roosta-Khorasani and M. W. Mahoney. Sub-sampled newton methods ii: Local convergence rates. *arXiv preprint arXiv:1601.04738*, 2016b.
- N. L. Roux, M. Schmidt, and F. R. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems*, pages 2663–2671, 2012.
- M. Schmidt, N. L. Roux, and F. R. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in neural information processing systems*, pages 1458–1466, 2011.
- S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb):567–599, 2013.
- S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, 155(1-2):105–145, 2016.
- O. Shamir, N. Srebro, and T. Zhang. Communication-efficient distributed optimization using an approximate Newton-type method. In *Proc. of the 31st Int. Conf. Machine Learning (ICML 2014)*, pages 1000–1008, 2014.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.
- S. Villa, S. Salzo, L. Baldassarre, and A. Verri. Accelerated and inexact forward-backward algorithms. *SIAM Journal on Optimization*, 23(3):1607–1633, 2013.
- J. Wang, W. Wang, and N. Srebro. Memory and communication efficient distributed stochastic optimization with minibatch prox. In *Proceedings of The 30th Conference on Learning Theory*, 2017a.
- X. Wang, S. Ma, D. Goldfarb, and W. Liu. Stochastic quasi-newton methods for nonconvex stochastic optimization. *SIAM Journal on Optimization*, 27(2):927–956, 2017b.
- B. E. Woodworth and N. Srebro. Tight complexity bounds for optimizing composite objectives. In *Advances in Neural Information Processing Systems*, pages 3639–3647, 2016.
- S. Wright and J. Nocedal. Numerical optimization. *Springer Science*, 35:67–68, 1999.
- L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.
- P. Xu, J. Yang, F. Roosta-Khorasani, C. Ré, and M. W. Mahoney. Sub-sampled newton methods with non-uniform sampling. In *Advances in Neural Information Processing Systems*, pages 3000–3008, 2016.
- T. Yang, R. Jin, S. Zhu, and Q. Lin. On data preconditioning for regularized loss minimization. *Machine Learning*, 103(1):57–79, 2016.

- Y. Zhang and L. Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 353–361, 2015.
- H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.