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

Jialei Wang jialei@uchicago.edu
University of Chicago
Tong Zhang tongzhang@tongzhang-ml.org
Tencent AI Lab

Abstract

We present a novel minibatch stochastic optimization method for empirical risk minimization problems. The method efficiently leverages variance reduced first-order and sub-sampled higher-order information to accelerate the convergence. We prove improved iteration complexity over state-of-the-art methods under suitable conditions.

1 Introduction

We consider the following optimization problem of finite-sums:

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

which arises in many machine learning problems. In the context of regularized loss minimization of linear predictors, we have $f_i(w) = \ell(w^\top x_i; b_i) + \frac{\lambda}{2}\|w\|^2$. Let $x_1, x_2, ..., x_n \in \mathbb{R}^d$ be feature vectors of n data samples, and $b_1, ..., b_n \in \mathbb{R}$ or $\{-1, 1\}$ be the corresponding target variables of interest. We note that (1) covers many popular models used in machine learning: for example when $\ell(w^\top x_i; b_i) = \log(1 + \exp(-b_i\langle w, x_i\rangle))$ we obtain ℓ_2 regularized logistic regression. The ubiquitousness of such finite-sum optimization problems and the massive scale of modern datasets motivate significant research effort on efficient optimization algorithms to solve (1). Throughout the paper, we assume each component function is L-smooth and λ-strongly convex.

In recent years there have been significant advances in developing fast optimization algorithms for (1), and we refer the readers to bottou2016optimization for a comprehensive survey of these developments. For large-scale problems in the form of (1), randomized methods are particularly efficient because of their low per iteration computation. Below we briefly review two lines of research: i) randomized variance reduced first-order methods; ii) randomized methods leveraging second-order information.

Variance Reduced First-order Methods The key technique for developing fast stochastic first-order methods is variance reduction, which makes the variance of the randomized update direction approching zero when the iterate gets closer to the optimum. Representative methods of this category include SAG roux2012stochastic, SVRG johnson2013accelerating, SDCA shalev2013stochastic and SAGA defazio2014saga, etc. In order to find a solution that reaches ϵ-suboptimality, these methods requires $\mathcal{O}\left(\left(n + \frac{L}{\lambda}\right)\log\left(\frac{1}{\epsilon}\right)\right)$ calls to the first-order oracle to compute the gradient of an individual function. In the large condition number regime (e.g. $\frac{L}{\lambda} > n$), by using an acceleration technique (such as Catalyst shalev2016accelerated,frostig2015regularizing,lin2015universal, APCG lin2015accelerated, SPDC zhang2015stochastic, Katyusha allen2016katyusha, etc), one can further improve the iteration complexity to $\mathcal{O}\left(\left(n + \sqrt{n\cdot\frac{L}{\lambda}}\right)\log\left(\frac{1}{\epsilon}\right)\right)$.

Leveraging Second-order Information Second-order information is 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, making vanilla Newton methods not well suited in solving (1). Therefore, there have been emerging studies in designing randomized algorithms that effectively utilize the *approximated* Hessian information. One line of such research is the sub-sampled Newton method which approximates the Hessian matrix based on a sub-sampled minibatch of data. Several work, such as byrd2011use,am2015sub,roosta2016sub,roosta2016sub2,bollapragada2016exact, established local linear

OPTML 2017: 10th NIPS Workshop on *Optimization for Machine Learning* (NIPS 2017).

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

Parameters η, b, $\widetilde{\lambda}$, ν, ε.
Initialize $\widetilde{w}_0 = 0$.
Sampling Sampling b items from $[n]$ to form a minibatch $\bar{B}$.
for $s = 1, 2, \ldots$ **do**
  Calculate $\widetilde{v} = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\widetilde{w}_{s-1})$.
  Initialize $y_0 = w_0 = \widetilde{w}_s$.
  for $t = 1, 2, \ldots, m$ **do**
    Sampling b items from $[n]$ to form a minibatch B_t.
    Find w_t **that approximately solve** (2) such that $\widetilde{f}_t(w_t) - \min_w \widetilde{f}_t(w) \leq \varepsilon$.

$$\widetilde{f}_t(w) := \frac{1}{2}(w - w_{t-1})^\top \left(\frac{1}{b}\sum_{i\in\bar{B}} \nabla^2 f_i(w_{t-1})\right)(w - w_{t-1}) + \left\langle \frac{\eta}{b}\sum_{i\in B_t} \nabla f_i(y_{t-1}) + \eta\widetilde{v} - \frac{\eta}{b}\sum_{i\in B_t} \nabla f_i(\widetilde{w}_{s-1}), w \right\rangle + \frac{\widetilde{\lambda}}{2}\|w - y_{t-1}\|^2. \tag{2}$$

    Update:

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

  end for
  Update $\widetilde{w}_s = w_m$.
end for
Return $\widetilde{w}_s$

convergence rates for different variants of sub-sampled Newton methods, when the size of the sub-sampled data is large enough. All of the aforementioned methods employ the full first-order gradient, and require calling the second-order oracle to compute the Hessian and solving a resulting linear system at every iteration. Therefore the computational complexity (as compared in Table 2 of xu2016sub) are often worse than stochastic algorithms such as SVRG. Another line of research is to consider randomness in both first and second-order information to design lower per-iteration cost algorithms. In particular, byrd2016stochastic considered combining minibatch SGD with Limited-memory BFGS (L-BFGS) liu1989limited type update. Inspired by this, moritz2016linearly,gower2016stochastic,wang2017stochastic proposed to combine variance reduced stochastic gradient with L-BFGS, and proved linear convergence for strongly convex and smooth objectives. However, theoretically it is hard to guarantee the quality of approximated Hessian using L-BFGS update, and thus the iteration complexity obtained by moritz2016linearly,gower2016stochastic,wang2017stochastic can be pessimistic, and can be much worse than vanilla SVRG.

In this paper, we make effort in this direction and make the following contributions: i) we propose 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 be naturally extended to solve composite optimization problems with non-smooth regularization; ii) we theoretically show under certain conditions the proposed approach can improve state-of-the-art iteration complexity, and empirically demonstrate it can substantially improve the convergence of existing methods.

2 Minibatch Stochastic Variance Reduced Proximal Iterations

In this section we present the proposed approach for minimizing (1). The high-level description is given in Algorithm 1, which consists of two main innovations: i) simultaneous incorporation of first-order and higher-order information via sub-sampling; ii) allowing larger minibatch sizes through acceleration with inexact minimization. We explain these features below.

Sampling both First-order and Higher-order Information

We first discuss the main building block of the algorithm. Suppose at iteration t, given the previous iterate w_{t-1}, we consider the following update rule:

$$w_t \approx \arg\min_w \frac{1}{2}(w - w_{t-1})^\top \left(\frac{1}{b}\sum_{i\in\bar{B}} \nabla^2 f_i(w_{t-1})\right)(w - w_{t-1}) + \frac{\widetilde{\lambda}}{2}\|w - w_{t-1}\|^2$$
$$+ \left\langle \frac{\eta}{b}\sum_{i\in B_t} \nabla f_i(w_{t-1}) + \eta\nabla f(\widetilde{w}) - \frac{\eta}{b}\sum_{i\in B_t}\nabla f_i(\widetilde{w}), w \right\rangle \quad (3)$$
$$= w_{t-1} - \eta\left(\frac{1}{b}\sum_{i\in\bar{B}}\nabla^2 f_i(w_{t-1}) + \widetilde{\lambda}I\right)^{-1}\left(\frac{1}{b}\sum_{i\in B_t}\nabla f_i(w_{t-1}) + \nabla f(\widetilde{w}) - \frac{1}{b}\sum_{i\in B_t}\nabla f_i(\widetilde{w})\right), \quad (4)$$

where $\bar{B}, B_t$ are some randomly sampled minibatch from $1, ..., n$, both with minibatch size b; η and $\widetilde{\lambda}$ are stepsize parameters, and $\widetilde{w}$ is a "reference" predictor used for reducing variance. The main feature of updating rule (3) is it considered both noisy first-order and higher-order information via minibatch sampling. The term $\frac{1}{b}\sum_{i\in\bar{B}} f_i(w) - \left\langle \frac{1}{b}\sum_{i\in\bar{B}}\nabla f_i(w_{t-1}), w\right\rangle$ in (3) incorporates noisy higher-order information of the objective function. It which can be treated as a variant of sub-sampled Newton method, combined with the minibatch stochastic gradient with variance reduction. And the update rule (3) can be treated as a preconditioned minibatch SVRG update rule, with $\left(\frac{1}{b}\sum_{i\in\bar{B}}\nabla^2 f_i(w_{t-1}) + \widetilde{\lambda}I\right)^{-1}$ as a precondition matrix.

Large Minibatch Size via Acceleration with Inexact Minimization

Using (3) as the building block, we propose the MB-SVRP (minibatch stochastic variance reduced proximal iterations) method, which is detailed in Algorithm 1. At the beginning of the algorithm, we form a minibatch $\bar{B}$ by sampling from $1, ..., n$ and fix it for the whole optimization process. Then following the SVRG method johnson2013accelerating, the algorithm is divided to multiple stages, indexed by s. At each stage, we iteratively solve a minimization problem of the form (2) based on the randomly sampled minibatch B_t.

Compared with the simple update rule in (3), the major difference in Algorithm 1 is that we consider a momentum scheme by maintaining two sequences $\{w_t, y_t\}$, which is inspired by Nesterov's acceleration technique nesterov2003introductory and its recent SVRG variant nitanda2014stochastic. The main theoretical advantage over minibatch SVRG without momentum konevcny2016mini is that such an acceleration allows us to use a much larger minibatch size (up to a size of $O(\sqrt{n})$) without slowing down the convergence.

Since it is often expensive to find the exact minimizer of (2), we consider an approximate minimizer with objective suboptimality ε. When we choose the appropriate $\widetilde{\lambda}$ for $\widetilde{f}_t(w)$ to obtain enough strong convexity, the subproblems can be efficiently solved to high accuracy efficiently. For example, using SVRG-type algorithms johnson2013accelerating to solve (2) allows us to find an approximate solution with a small suboptimality ε with a constant number of passes over the data in the minibatch $\bar{B} \cup B_t$, if $\widetilde{\lambda}$ is set appropriately.

3 Theoretical Results

Besides the strong convexity and smoothness, we also need the following Hessian Lipschitz condition, and the notion of effective dimension and statistical leverage.

Condition 1. *For each component function $f_i(w), \forall i \in [n]$, its Hessian matrix is M-Lipschitz, i.e.*

$$\left\|\nabla^2 f_i(w) - \nabla^2 f_i(w')\right\| \le M\|w - w'\|, \forall w, w' \in \mathbb{R}^d, \forall i \in [n].$$

Definition 1. ***(Effective dimension)*** *Let the $\lambda_1, ..., \lambda_d$ be the top-d eigenvalues of $H_0(w^*) = (1/n)\sum_{i=1}^n \ell''({w^*}^\top x_i; b_i)x_i x_i^\top$, define the effective dimension $d_{\widetilde{\lambda}}$ (for some $\widetilde{\lambda} \ge 0$) of $H_0(w^*)$ being $d_{\widetilde{\lambda}} = \sum_{j=1}^d \frac{\lambda_j}{\lambda_j + \widetilde{\lambda}}$.*

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

Definition 2. *(**Statistical leverage at** $\widetilde{\lambda}$) Let* $H_{\widetilde{\lambda}}(w^*) = \frac{1}{n}\sum_{i=1}^n \ell''(w^{*\top} x_i; b_i) x_i x_i^\top + \widetilde{\lambda} I$, *we say the statistical leverage of data matrix* X *is bounded by* $\rho_{\widetilde{\lambda}}$ *at* $\widetilde{\lambda}$ *if* $\max_{i\in[n]} \frac{\left\| H_{\widetilde{\lambda}}^{-1/2}(w^*)\ell''(w^{*\top}x_i;b_i)^{1/2} x_i \right\|}{\sqrt{(1/n)\sum_{j=1}^n \left\| H_{\widetilde{\lambda}}^{-1/2}(w^*)\ell''(w^{*\top}x_j;b_j)^{1/2} x_j \right\|^2}} \leq \rho_{\widetilde{\lambda}}$.

Theorem 3. *Consider Algorithm 1 on problems with L-smooth and λ-strongly convex functions that satisfied Condition 1 with Hessian Lipschitz parameter M. Suppose we sample minibatch $\bar{B}$ uniformly from $1, ..., n$, and set the tuning parameters as*

$$\widetilde{\lambda} = \max\left\{\lambda, \frac{L}{b}\right\}, \quad b \asymp \min\left\{n, \rho_{\widetilde{\lambda}}^2 d_{\widetilde{\lambda}} \left(\frac{L}{\lambda}\right)^{1/3}\right\}, \quad \eta \asymp \min\left\{\frac{b^3\lambda}{(\rho_{\widetilde{\lambda}}^2 d_{\widetilde{\lambda}})^2 L}, \frac{1}{\rho_{\widetilde{\lambda}}^2 d_{\widetilde{\lambda}}}\right\},$$

and suppose we start from the initialization point that is sufficiently close to the optimum: $\|w_0 - w^*\| \leq R = \mathcal{O}\left(\frac{\lambda^4}{L^2 M}\right)$. Given $\epsilon > 0$, and let $\varepsilon \leq \frac{1}{10^5} \cdot \left(\frac{\lambda}{L}\right)^7 \epsilon$. Then for the MP-SVRP algorithm to find the approximate solution $\widetilde{w}_s$ of (1) that reaches the expected ϵ-objective suboptimality $\mathbb{E} f(\widetilde{w}_s) - \min_w f(w) \leq \epsilon$, the total number of oracle calls to solve (2) is no more than

$$\mathcal{O}\left(\max\left\{\left(\frac{L}{\lambda}\right)^{1/3}, \frac{\rho_{\widetilde{\lambda}}^2 d_{\widetilde{\lambda}} L}{\lambda n^2}\right\} \cdot \log\left(\frac{1}{\epsilon}\right)\right).$$

Note that in Algorithm 1, we can use SVRG to solve each subproblem in (2). This leads to the following result.

Corollary 4. *Assume that the conditions of Theorem 3 hold. Moreover, if we use SVRG to solve each subproblem in (2) up to the suboptimality ε, then the total number of gradient evaluations used in the whole MP-SVRP algorithm can be upper bounded by*

$$\mathcal{O}\left(\max\left\{\rho_{\widetilde{\lambda}}^2 d_{\widetilde{\lambda}} \left(\frac{L}{\lambda}\right)^{2/3}, \frac{\rho_{\widetilde{\lambda}}^2 d_{\widetilde{\lambda}} L}{\lambda n}\right\} \cdot \log\left(\frac{L}{\lambda}\right) \cdot \log^2\left(\frac{1}{\epsilon}\right) + n \cdot \log\left(\frac{1}{\epsilon}\right)\right).$$

Acceleration Corollary 4 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 shalev2016accelerated, frostig2015regularizing,lin2015universal, it is possible to obtain an accelerated convergence rate which has milder dependence on condition number, Table 1 summarized the main results.

References

[Allen-Zhu(2016)] Z. Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *arXiv preprint arXiv:1603.05953*, 2016.

[Bollapragada et al.(2016)Bollapragada, Byrd, and Nocedal] R. Bollapragada, R. Byrd, and J. Nocedal. Exact and inexact subsampled newton methods for optimization. *arXiv preprint arXiv:1609.08502*, 2016.

[Bottou et al.(2016)Bottou, Curtis, and Nocedal] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *arXiv preprint arXiv:1606.04838*, 2016.

[Byrd et al.(2011)Byrd, Chin, Neveitt, and Nocedal] 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.

[Byrd et al.(2016)Byrd, Hansen, Nocedal, and Singer] 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.

[Cotter et al.(2011)Cotter, Shamir, Srebro, and Sridharan] 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.

[Defazio et al.(2014)Defazio, Bach, and Lacoste-Julien] 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.

[Dekel et al.(2012)Dekel, Gilad-Bachrach, Shamir, and Xiao] 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.

[Devolder et al.(2014)Devolder, Glineur, and Nesterov] 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.

[Erdogdu and Montanari(2015)] M. A. Erdogdu and A. Montanari. Convergence rates of sub-sampled newton methods. In *Advances in Neural Information Processing Systems*, pages 3052–3060, 2015.

[Frostig et al.(2015)Frostig, Ge, Kakade, and Sidford] R. Frostig, R. Ge, S. Kakade, and A. Sidford. Unregularizing: 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.

[Gower et al.(2016)Gower, Goldfarb, and Richtárik] 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.

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

[Konečnỳ et al.(2016)Konečnỳ, Liu, Richtárik, and Takáč] J. Konečnỳ, J. Liu, P. Richtárik, and M. Takáč. Mini-batch semi-stochastic gradient descent in the proximal setting. *IEEE Journal of Selected Topics in Signal Processing*, 10(2):242–255, 2016.

[Lan(2012)] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

[Lin et al.(2015a)Lin, Mairal, and Harchaoui] 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.

[Lin et al.(2015b)Lin, Lu, and Xiao] 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.

[Liu and Nocedal(1989)] D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

[Moritz et al.(2016)Moritz, Nishihara, and Jordan] P. Moritz, R. Nishihara, and M. Jordan. A linearly-convergent stochastic l-bfgs algorithm. In *Artificial Intelligence and Statistics*, pages 249–258, 2016.

[Nesterov(2004)] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2004.

[Nitanda(2014)] A. Nitanda. Stochastic proximal gradient descent with acceleration techniques. In *Advances in Neural Information Processing Systems*, pages 1574–1582, 2014.

[Roosta-Khorasani and Mahoney(2016a)] F. Roosta-Khorasani and M. W. Mahoney. Sub-sampled newton methods i: globally convergent algorithms. *arXiv preprint arXiv:1601.04737*, 2016a.

[Roosta-Khorasani and Mahoney(2016b)] F. Roosta-Khorasani and M. W. Mahoney. Sub-sampled newton methods ii: Local convergence rates. *arXiv preprint arXiv:1601.04738*, 2016b.

[Roux et al.(2012)Roux, Schmidt, and Bach] 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.

[Schmidt et al.(2011)Schmidt, Roux, and Bach] 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.

[Shalev-Shwartz and Zhang(2013)] 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.

[Shalev-Shwartz and Zhang(2016)] S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, 155(1-2):105–145, 2016.

[Wang et al.(2017a)Wang, Wang, and Srebro] 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.

[Wang et al.(2017b)Wang, Ma, Goldfarb, and Liu] 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.

[Xu et al.(2016)Xu, Yang, Roosta-Khorasani, Ré, and Mahoney] 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.

[Zhang and Xiao(2015)] 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.