CSE 599: Interplay between Convex Optimization and Geometry **Winter 2018**

Lecture 8: Stochastic Methods and Applications

Lecturer: Yin Tat Lee

Disclaimer: *Please tell me any mistake you noticed.*

Last lecture, we showed that accelerated coordinate descent takes $O^*(\sqrt{d \cdot \frac{\sum_i \beta_i}{\alpha}})$ for α strongly convex function on $\mathbb{R}^d$ with $\frac{\partial^2}{\partial x_i^2} f(x) \leq \beta_i$. This can be improved to $O^*(\sum_i \sqrt{\frac{\beta_i}{\alpha}})$ [47]. It can be proved using the linear coupling techniques. It seems to me that one cannot recover such result using the reduction techniques.

8.1 (Accelerated) Stochastic Methods

8.1.1 Motivation

Suppose the function we want to minimize is given by sum of many convex functions (namely, $\ell = \frac{1}{n}\sum_{i=1}^n \ell_i$). It makes sense to make a gradient update using only one term, especially if ℓ_i are similar to each other. Naively, one may sample a term ℓ_i randomly and update the solution by the gradient step

$$x^{(k+1)} = x^{(k)} - \eta \nabla \ell_i(x^{(k)}). \tag{8.1}$$

We can think $\eta \nabla \ell_i(x^{(k)})$ is an unbiased estimator $\eta \nabla \ell(x^{(k)})$. Therefore, if η is small enough, this algorithm is essentially same as the gradient descent. To analyze such algorithm, we need to estimate the variance of this estimator. However, we note that $\nabla \ell_i(x^*)$ can be very large even at the optimal x^*. Therefore, the variance of the estimator can be huge even if we are close to the solution.

As an example, consider the simplest 1 dimension problem $\ell(x) = \frac{1}{n}\sum_{i=1}^n (x - b_i)^2$. Let suppose that $|b_i| \leq 1$. It is easy to see the minimizer is $x^* = \frac{1}{n}\sum_{i=1}^n b_i$ and that $x^{(T)} = x^* + O(\frac{1}{\sqrt{T}})$. Therefore, the algorithm (8.1) only gives $\ell(x^{(T)}) - \min_x \ell(x) = O(\frac{1}{T})$ for this simple problem. This type of guarantee is optimal for the case $n = \infty$. However, if we can read all the input, it is possible to avoid this variance problem.

8.1.2 Variance Reduction

One can reduce the variance of the estimator $\eta \nabla \ell_i(x^{(k)})$ as follows

$$x^{(k+1)} = x^{(k)} - \eta(\nabla \ell_i(x^{(k)}) - \nabla \ell_i(x^{(0)}) + \nabla \ell(x^{(0)})) \tag{8.2}$$

Note that again we update x with an unbiased estimator of $\eta \nabla \ell(x^{(k)})$. However, this estimator has much smaller variance. When x is close to x^*, $\nabla \ell(x)$ is close to 0. Formally, we can bound the difference of gradient by the following lemma:

Lemma 8.1.1. *For any β smooth convex function f, we have that*

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2\beta} \|\nabla f(y) - \nabla f(x)\|_2^2 .$$

Proof. Let $g(y) = f(y) - f(x) - \nabla f(x)^\top (y - x)$. By convexity of f, $g \geq 0$. Since g is β smooth, we have

$$0 \leq g(y - \frac{1}{\beta}\nabla g(y)) \leq g(y) - \frac{1}{2\beta}\left\|\nabla g(y)\right\|_2^2.$$

Using $\left\|\nabla g(y)\right\|_2 = \left\|\nabla f(y) - \nabla f(x)\right\|_2$, we get the result. □

Now, we can bound the variance of the estimator used in (8.2).

Lemma 8.1.2. *Given a convex function $\ell = \frac{1}{n}\sum \ell_i$. Suppose that ℓ_i is β_i-smooth and let $\beta = \frac{1}{n}\sum \beta_i$. If we sample term i with probability $p_i = \frac{\beta_i}{n\beta}$ and let*

$$h_i = \frac{1}{\beta_i}\nabla \ell_i(x) - \frac{1}{\beta_i}\nabla \ell_i(x^{(0)}) + \frac{1}{\beta}\nabla \ell(x^{(0)}).$$

Then, we have that

$$\mathbb{E}_i h_i = \frac{1}{\beta}\nabla \ell(x), \quad \mathbb{E}_i \left\|h_i\right\|_2^2 = \frac{4}{\beta}(\ell(x) - \ell(x^*)) + \frac{4}{\beta}(\ell(x^{(0)}) - \ell(x^*)).$$

Proof. The mean $\mathbb{E}_i h_i$ follows from simple calculations. For the variance, we have

$$\begin{aligned}\mathbb{E}_i \left\|h_i\right\|_2^2 &= \sum_i \frac{\beta_i}{n\beta}\left\|\frac{1}{\beta_i}\nabla \ell_i(x) - \frac{1}{\beta_i}\nabla \ell_i(x^{(0)}) + \frac{1}{\beta}\nabla \ell(x^{(0)})\right\|_2^2 \\ &\leq \sum_i \frac{2}{n\beta\beta_i}\left\|\nabla \ell_i(x) - \nabla \ell_i(x^*)\right\|_2^2 + \sum_i \frac{2\beta_i}{n\beta}\left\|\frac{1}{\beta}\nabla \ell(x^{(0)}) - \frac{1}{\beta_i}\nabla \ell_i(x^{(0)}) + \frac{1}{\beta_i}\nabla \ell_i(x^*)\right\|_2^2 \end{aligned} \tag{8.3}$$

For the first term, Lemma 8.1.1 shows that

$$\sum_i \frac{1}{\beta_i}\left\|\nabla \ell_i(x) - \nabla \ell_i(x^*)\right\|_2^2 \leq 2\sum_i (\ell_i(x) - \ell_i(x^*) - \nabla \ell_i(x^*)^\top (x - x^*)) \leq 2n(\ell(x) - \ell(x^*)). \tag{8.4}$$

For the second term, we note that

$$\mathbb{E}(\frac{1}{\beta_i}\nabla \ell_i(x^{(0)}) - \frac{1}{\beta_i}\nabla \ell_i(x^*)) = \frac{1}{\beta}\nabla \ell(x^{(0)}) - \frac{1}{\beta}\nabla \ell(x^*) = \frac{1}{\beta}\nabla \ell(x^{(0)}).$$

Using $\mathbb{E}\left\|X - \mathbb{E}X\right\|_2^2 \leq \mathbb{E}\left\|X\right\|_2^2$, we have

$$\begin{aligned}\sum_i \beta_i \left\|\frac{1}{\beta}\nabla \ell(x^{(0)}) - \frac{1}{\beta_i}\nabla \ell_i(x^{(0)}) - \frac{1}{\beta_i}\nabla \ell_i(x^*)\right\|_2^2 &\leq \sum_i \frac{1}{\beta_i}\left\|\frac{1}{\beta_i}\nabla \ell_i(x^{(0)}) - \frac{1}{\beta_i}\nabla \ell_i(x^*)\right\|_2^2 \\ &\leq 2n(\ell(x^{(0)}) - \ell(x^*))\end{aligned} \tag{8.5}$$

where the last sentence follows from the same reasoning in (8.4).

Combining (8.3), (8.4) and (8.5), we have that

$$\mathbb{E}_i \left\|h_i\right\|_2^2 \leq \frac{4}{\beta}(\ell(x) - \ell(x^*)) + \frac{4}{\beta}(\ell(x^{(0)}) - \ell(x^*)).$$

□

8.1.3 Analysis

Now, we are already to prove the progress of stochastic method for each step.

Lemma 8.1.3. *Given a convex function $\ell = \frac{1}{n}\sum \ell_i$. Suppose that ℓ_i is β_i-smooth and let $\beta = \frac{1}{n}\sum \beta_i$ and that ℓ is α strongly convex. Consider the algorithm $x^{(k+1)} \leftarrow x^{(k)} - \eta \cdot h_i^{(k)}$ where*

$$h_i^{(k)} = \frac{1}{\beta_i}\nabla \ell_i(x^{(k)}) - \frac{1}{\beta_i}\nabla \ell_i(x^{(0)}) + \frac{1}{\beta}\nabla \ell(x^{(0)}).$$

For $\eta \le \frac{1}{4}$, we have

$$\mathbb{E}\ell(\frac{1}{T}\sum_{k=0}^{T-1} x^{(k)}) - \ell(x^*) \le (\frac{2\beta}{\alpha\eta} + 4\eta T)(\ell(x^{(0)}) - \ell(x^*)).$$

Setting $\eta = \Theta(\sqrt{\frac{\beta}{\alpha T}})$ and $T = \Theta(\frac{\beta}{\alpha})$, we get that

$$\mathbb{E}\ell(\frac{1}{T}\sum_{k=0}^{T-1} x^{(k)}) - \ell(x^*) \le \frac{1}{2}(\ell(x^{(0)}) - \ell(x^*)).$$

Proof. Note that

$$\mathbb{E}_i \left\|x^{(k+1)} - x^*\right\|_2^2 = \left\|x^{(k)} - x^*\right\|_2^2 - 2\eta \cdot \mathbb{E}_i h_i^{(k)\top}(x^{(k)} - x^*) + \eta^2 \mathbb{E}_i \left\|h_i^{(k)}\right\|_2^2. \tag{8.6}$$

Using the formula of $\mathbb{E}_i h_i^{(k)}$ and $\mathbb{E}_i \left\|h_i^{(k)}\right\|_2^2$ (Lemma 8.1.2), we have that

$$\begin{aligned}
\mathbb{E}_i \left\|x^{(k+1)} - x^*\right\|_2^2 &\le \left\|x^{(k)} - x^*\right\|_2^2 - \frac{2\eta}{\beta}\cdot \nabla \ell(x^{(k)})^\top (x^{(k)} - x^*) + \frac{4\eta^2}{\beta}(\ell(x^{(k)}) + \ell(x^{(0)}) - 2\ell(x^*)) \\
&\le \left\|x^{(k)} - x^*\right\|_2^2 - \frac{2\eta}{\beta}\cdot (\ell(x^{(k)}) - \ell(x^*)) + \frac{4\eta^2}{\beta}(\ell(x^{(k)}) + \ell(x^{(0)}) - 2\ell(x^*)) \\
&\le \left\|x^{(k)} - x^*\right\|_2^2 - \frac{\eta}{\beta}\cdot (\ell(x^{(k)}) - \ell(x^*)) + \frac{4\eta^2}{\beta}(\ell(x^{(0)}) - \ell(x^*))
\end{aligned}$$

where we used that $\eta \le \frac{1}{4}$. Summing up T terms, we have

$$\begin{aligned}
\mathbb{E}\left\|x^{(T)} - x^*\right\|_2^2 &\le \left\|x^{(0)} - x^*\right\|_2^2 - \frac{\eta}{\beta}\cdot \mathbb{E}\sum_{k=0}^{T-1}(\ell(x^{(k)}) - \ell(x^*)) + \frac{4\eta^2 T}{\beta}(\ell(x^{(0)}) - \ell(x^*)) \\
&\le (\frac{2}{\alpha} + \frac{4\eta^2 T}{\beta})(\ell(x^{(0)}) - \ell(x^*)) - \frac{\eta}{\beta}\cdot \mathbb{E}\sum_{k=0}^{T-1}(\ell(x^{(k)}) - \ell(x^*)).
\end{aligned}$$

In particular, we have that

$$\begin{aligned}
\mathbb{E}\ell(\frac{1}{T}\sum_{k=0}^{T-1} x^{(k)}) - \ell(x^*) &\le \frac{\beta}{\eta T}(\frac{2}{\alpha} + \frac{4\eta^2 T}{\beta})(\ell(x^{(0)}) - \ell(x^*)) \\
&= (\frac{2\beta}{\alpha\eta T} + 4\eta)(\ell(x^{(0)}) - \ell(x^*)).
\end{aligned}$$

□

By restarting the algorithm above, we have the following result:

Theorem 8.1.4. *Given a convex function $\ell = \frac{1}{n}\sum \ell_i$. Suppose that ℓ_i is β_i-smooth and let $\beta = \frac{1}{n}\sum \beta_i$ and that ℓ is α strongly convex. Suppose we can compute $\nabla \ell_i$ in time $\mathcal{T}_{\text{stoc}}$ and $\nabla \ell$ in time $\mathcal{T}_{\text{full}}$. We have an algorithm that outputs an x such that*

$$\mathbb{E}\ell(x) - \ell(x^*) \le \varepsilon(\ell(x^{(0)}) - \ell(x^*))$$

in time $O((\mathcal{T}_{\text{full}} + \kappa \mathcal{T}_{\text{stoc}}) \log(\frac{1}{\varepsilon}))$ with $\kappa = \frac{\beta}{\alpha}$.

Similar to the coordinate descent, we can accelerate it using the accelerated gradient descent (Theorem 7.2.1). To apply Theorem 7.2.1, we consider the function

$$\phi(x) = f(x) + h(x) \quad \text{with} \quad f(x) = \frac{\alpha}{2}\|x\|_2^2 \text{ and } h(x) = \ell(x) - \frac{\alpha}{2}\|x\|_2^2.$$

For simplicity, we assume $\mathcal{T}_{\text{full}} = n$ and $\mathcal{T}_{\text{stoc}} = 1$. Since f is $\alpha + \frac{\beta}{n}$ smooth (YES!, I know this is also α smooth) and α strongly convex and since h is convex, we apply Theorem 7.4.1 and get an algorithm that takes $O^*(1 + \sqrt{\frac{\kappa}{n}})$ steps. Note that each step involves $\mathcal{T}_f + \mathcal{T}_{h,\alpha+\frac{\beta}{n}}$. Obviously, $\mathcal{T}_f = 0$. Next, note that $\mathcal{T}_{h,\alpha+\frac{\beta}{n}}$ involves solving a problem of the form

$$\begin{aligned} y_x &= \text{argmin}_y (\frac{\alpha}{2} + \frac{\beta}{2n})\|y - x\|^2 + (\ell(y) - \frac{\alpha}{2}\|x\|^2) \\ &= \text{argmin}_y \ell(y) - \alpha y^\top x + \frac{\beta}{2n}\|y - x\|^2 \\ &= \text{argmin}_y \frac{1}{n}\sum_i (\ell_i(y) + \frac{\beta}{2n}\|y - x\|^2 - \alpha y^\top x) \end{aligned}$$

Now, we can apply Theorem 8.1.4 to solve this problem. It takes

$$O^*(n + \frac{\beta + \frac{\beta}{n}}{\alpha + \frac{\beta}{n}}) = O^*(n)$$

Therefore, in total it takes

$$O^*(1 + \sqrt{\frac{\kappa}{n}}) \cdot O^*(n) = O^*(n + \sqrt{\kappa n}).$$

Theorem 8.1.5. *Given a convex function $\ell = \frac{1}{n}\sum \ell_i$. Suppose that ℓ_i is β_i-smooth and let $\beta = \frac{1}{n}\sum \beta_i$ and that ℓ is α strongly convex. Suppose we can compute $\nabla \ell_i$ in $O(1)$ time. We have an algorithm that outputs an x such that*

$$\mathbb{E}\ell(x) - \ell(x^*) \le \varepsilon(\ell(x^{(0)}) - \ell(x^*))$$

in time $O^*(n + \sqrt{\kappa n})$ with $\kappa = \frac{\beta}{\alpha}$.

In general, the lower bound is $\Omega^*(n + \sqrt{\kappa d})$ (Note that that paper claimed $n + \sqrt{\kappa n}$ but their lower bound instances has $n \sim d$) [49]. For quadratic case, $O^*(n + \sqrt{\kappa d})$ is indeed possible [45]. This improvement from n to d can be significant as n (the number of samples) is in some cases orders of magnitude larger than d (the number of features). For example, in the LIBSVM dataset, in 87 out of 106 many non-text problems, we have $n \ge d$, 50 of them have $n \ge d^2$ and in the UCI dataset, in 279 out of 301 many non-text problems, we have $n \ge d$, 195 out of them have $n \ge d^2$.

Problem 8.1.6. Is $O^*(n + \sqrt{\kappa d})$ possible?

8.1.4 Relation between Coordinate Descent and Stochastic Method

Consider the problem $\min_x \sum_{i=1}^n \ell_i(a_i^\top x) + \frac{\alpha}{2}\|x\|^2$. Let A be the matrix with rows given by a_i. One can compute its dual as follows

$$\begin{aligned}\min_x \sum_{i=1}^n \ell_i(a_i^\top x) + \frac{\alpha}{2}\|x\|^2 &= \min_x \max_\theta \theta^\top Ax - \sum_{i=1}^n \ell_i^*(\theta_i) + \frac{\alpha}{2}\|x\|^2 \\ &= \max_\theta \min_x x^\top A^\top \theta - \sum_{i=1}^n \ell_i^*(\theta_i) + \frac{\alpha}{2}\|x\|^2 \\ &= \max_\theta -\sum_{i=1}^n \ell_i^*(\theta_i) - \frac{1}{2\alpha}\left\|A^\top\theta\right\|^2\end{aligned}$$

where we used that $x = -\frac{1}{\alpha}A^\top\theta$.

Note that the stochastic method involves (consider the version without variance reduction for simplicity)

$$x^{(k+1)} = x^{(k)} - \eta \cdot \nabla \ell_i(a_i^\top x) \cdot a_i - \eta \cdot \alpha x^{(k)} \tag{8.7}$$

and the coordinate descent on the dual involves

$$\theta_i^{(k+1)} = \theta_i^{(k)} + \eta \cdot \nabla \ell_i^*(\theta_i^{(k)}) + \frac{1}{\alpha}(AA^\top\theta^{(k)})_i.$$

Let $\widehat{x} = -\frac{1}{\alpha}A^\top\theta$ (used the hat to emphasis this comes from the coordinate descent), we have that

$$\widehat{x}^{(k+1)} = \widehat{x}^{(k)} - \frac{\eta}{\alpha} \cdot \nabla \ell_i^*(\theta_i) \cdot a_i + \frac{1}{\alpha}(a_i^\top \widehat{x}) \cdot a_i. \tag{8.8}$$

Although (8.7) and (8.8) is not exactly the same, they have some similarity. They update the step only using one ℓ_i and only move on the direction a_i. In this setting, in fact, one can recover the guarantee of stochastic descent (with variance reduction) via coordinate descent.

Problem 8.1.7. Since we can solve problem in $O^*(\sum_i \sqrt{\beta_i/\alpha})$ using coordinate descent, is it possible to achieve $O^*(\sum_i \sqrt{\beta_i/\alpha})$ for stochastic method. (Note that the β_i in coordinate descent is different from the β_i in stochastic method.

8.1.5 Discussion on Linear systems

Note that when we used accelerated gradient descent on ℓ directly, the running time is $O^*(\sqrt{\kappa}n)$. Therefore, this is a $\sqrt{n}$ factor improvement in general. To better understand the guarantee of stochastic descent, one can consider the simplest setting

$$\ell(x) = \sum_{i=1}^n (a_i^\top x - b_i)^2.$$

For simplicity, we assume that each row of a_i has $O(1)$ non-zeros and hence each step of stochastic step takes $O(1)$ time. Applying Theorem 8.1.5, we have that

Corollary 8.1.8. *Given a matrix $A \in \mathbb{R}^{n\times d}$. Assume each row of A has $O(1)$ non-zeros. We can find a random x such that*

$$\mathbb{E}\|Ax-b\|_2^2 - \min_x \|Ax-b\|_2^2 \le \varepsilon\left(\|b\|_2^2 - \min_x \|Ax-b\|_2^2\right)$$

in time $O^*(n + \sqrt{\frac{n\cdot\|A\|_F^2}{\lambda_{\min}}})$ where $\lambda_{\min}$ is the minimum eigenvalue of $A^\top A$.

Proof. Note that $(a_i^\top x - b_i)^2$ is $\|a_i\|_2^2$ smooth. Therefore, the smoothness of the problem is $\beta = \sum_i \|a_i\|_2^2 = \|A\|_F^2$. Clearly, the strong convexity of the problem is $\lambda_{\min}$. This gives the result. □

As a comparison, we note that accelerated gradient descent takes $O^*(n \cdot \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}})$ time. Note that

$$\sqrt{\frac{n \cdot \|A\|_F^2}{\lambda_{\min}}} = \sqrt{\frac{n \cdot \sum_i \lambda_i}{\lambda_{\min}}} \leq n \cdot \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}.$$

So, the accelerated stochastic method is better than accelerated gradient descent. Furthermore, we see that it is much better if there is only few large eigenvalues. Unfounrtately, when the eigenvalue distributed uniformly, then this does not give any improvement. For example, this does not give an improvement for the lower bound instance $\sum_i (x_i - x_{i+1})^2$ we mentioned many times.

8.1.6 Limitation of Reduction

In this and last lecture, we demonstrate how to obtain accelerated methods via a reduction. However, they have several limitation. First of all, the algorithm is not as clean as direct method and the running time has some extra logs. Sometimes, it is better to sample few terms ℓ_i at the same time. This allows you to group some computation together and hence lower cost per term. We call the number of terms we used each iteration is the batch size. It is natural to ask what is the largest batch size we can use per iteration while not affecting the convergence rate.

Theorem 8.1.9 ([46]). *Given a convex function $\ell = \frac{1}{n}\sum \ell_i$. Suppose that ℓ_i is β_i-smooth and that ℓ is B-smooth. Then, we can take batch size $\sqrt{\frac{n \cdot \sum \beta_i}{B}}$ without affecting the convergence rate by more than a constant.*

8.1.7 Application to Laplacian systems

Given a graph $G = (V, E)$ with m edges and n vertices. We consider the problem

$$\frac{1}{2} \sum_{(i,j) \sim E} (x_i - x_j)^2 - c^\top x.$$

For now, we can think this as a generalization of the worst case function. Directly applying accelerated gradient descent, accelerated coordinate descent or accelerated stochastic descent all gives $O^*(mn)$ time algorithm. Roughly speaking, this is because of the "diameter" bottleneck for all first-order methods.

To get around the diameter issue, we let the incidence matrix $B \in \mathbb{R}^{|E| \times |V|}$ defined by $B_{(i,j),i} = 1$ and $B_{(i,j),j} = -1$.We consider the dual problem

$$\min_{B^T f = d} \frac{1}{2} \sum_{e \in E} f_e^2.$$

Let pick a tree T, by sending flow along the tree, it is easy to find g such that $B^\top g = d$. Hence, the problem can be rewritten as

$$\min_{B^T f = 0} \frac{1}{2} \sum_{e \in E} (f + g)_e^2.$$

We call any f satisfying $B^\top f = 0$ is a circulation. For any edge $e \notin T$, we define $c(e)$ be an unit cycle on $\{e\} \cup T$. It is known that any circulation can be represented by

$$f = \sum_{e \notin T} \alpha_e c(e).$$

Also, any such flow is a circulation. Therefore, the problem now becomes

$$\min_{\alpha} \ell(\alpha) \stackrel{\text{def}}{=} \frac{1}{2} \sum_{e \in E} (\sum_{k \notin T} \alpha_k c(k) + g_e)_e^2.$$

Now, we apply accelerated coordinate descent to solve this problem. First, we note that ℓ is 1 strongly convex. Next, for any k, $\frac{\partial^2}{\partial \alpha_k^2} \ell = \text{length}(c(k))$. Hence, the accelerated coordinate descent takes $O^*(\sum_{k \notin T} \text{length}(c(k)))$ coordinate steps. It is known that for any graph G, we can find a tree such that $\sum_{k \notin T} \text{length}(c(k)) = O^*(m)$. Using such tree, the running time is $O^*(m)$ coordinate steps. It is also easy to implement the algorithm such that each step takes $O^*(1)$ time. Therefore, this gives a $O^*(m)$ algorithm for the problem. See [48] for more details.

References

[45] Naman Agarwal, Sham Kakade, Rahul Kidambi, Yin Tat Lee, Praneeth Netrapalli, and Aaron Sidford. Leverage score sampling for faster accelerated regression and erm. *arXiv preprint arXiv:1711.08426*, 2017.

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

[47] Zeyuan Allen-Zhu, Zheng Qu, Peter Richtárik, and Yang Yuan. Even faster accelerated coordinate descent using non-uniform sampling. In *International Conference on Machine Learning*, pages 1110–1119, 2016.

[48] Yin Tat Lee and Aaron Sidford. Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 147–156. IEEE, 2013.

[49] Blake E Woodworth and Nati Srebro. Tight complexity bounds for optimizing composite objectives. In *Advances in neural information processing systems*, pages 3639–3647, 2016.