EE C227C Convex Optimization and Approximation Lecture 26 - 4/21/2022

Lecture 26: Geometrization and Randomized Coordinate Descent

Lecturer: Jiantao Jiao *Scribe: Sahil Patel, Hanlin Zhu*

1 Review

We begin with a brief review of the various stochastic optimization methods we have discussed thus far. In particular, we consider the specific problem formulation where our objective function is given by the sum of a set of functions,

$$\min_x f(x) = \min_x \frac{1}{n}\sum_{i=1}^n f_i(x).$$

We assume that evaluating $\nabla f_i(x)$ takes $O(1)$ time, which implies that evaluating $\nabla f(x) = \frac{1}{n}\sum_{i=1}^n \nabla f_i(x)$ takes $O(n)$ time. Further, we assume that each f_i is L-smooth and that f is μ-strongly convex. For gradient descent (GD), we recall that the number of iterations to achieve ϵ accuracy is $O(\kappa \log(1/\epsilon))$, where we define the condition number $\kappa = L/\mu$, and given that the entire gradient $\nabla f(x)$ must be computed at every iteration, the overall query complexity is $O(n\kappa \log(1/\epsilon))$. To attain the same accuracy, stochastic gradient descent (SGD) requires a greater number of iterations, $O(1/(\mu\epsilon))$, due to the effect of stochasticity, however, the cost of of each iteration is only $O(1)$ to evaluate a single gradient $\nabla f_i(x)$ such that the query complexity is $O(1/(\mu\epsilon))$. Stochastic variance reduced gradient (SVRG) achieves a query complexity of $O((n+\kappa)\log(1/\epsilon))$, which can be seen as a sort of "best of both worlds" between GD and SGD as we now have a factor of $n+\kappa$ instead of $n\kappa$ and the asymptotic behavior of $\log(1/\epsilon)$ is better than $1/\epsilon$ for small ϵ.

2 Stochastic Variance Reduced Gradient

2.1 Geometrization

We recall from the last lecture that we proved the following performance guarantee for SVRG:

Theorem 1. *For $s \geq 0$, we have*

$$2\eta(1-2\eta L)\left(\mathbb{E}f\left(\bar{x}_{s+1}\right) - f^*\right) + \eta\mu\mathbb{E}\left\|\tilde{x}_{s+1} - x^*\right\|^2 \leq 4L\eta^2\left(\mathbb{E}f\left(\tilde{x}_s\right) - f^*\right) + \frac{1}{m}\mathbb{E}\left\|\bar{x}_s - x^*\right\|^2$$

If $\eta = 1/(8L)$ and $m \geq 2/(\eta\mu) = 16L/\mu = 16\kappa$, then

$$\mathbb{E}f\left(\tilde{x}_{s+1}\right) - f^* + \mu\mathbb{E}\left\|\tilde{x}_{s+1} - x^*\right\|^2 \leq \left(\frac{1}{2}\right)^s\left[f\left(\tilde{x}_0\right) - f^* + \mu\left\|\tilde{x}_0 - x^*\right\|^2\right]$$

We note that to achieve the exponential decay in our particular measure of suboptimality, which combines the absolute suboptimality with the distance of the final iterate to the minimum, as described in the final line of the theorem, the upper bound on the length of each epoch m must depend on the condition number $\kappa = \mu/L$ and by extension the strong convexity parameter μ, which may be difficult to estimate, especially compared to the smoothness parameter L. Therefore, it would be desirable to avoid such a dependence, and the technique of geometrization provides a means of precisely accomplishing this [1]. Specifically, we consider the randomly chosen length of each epoch N to now be distributed geometrically with respect to m rather than uniformly as before. We also leverage a special property which can be shown to only hold for geometric distributions,

Lemma 2. *Let $N \sim Geom(m)$ for $m > 0$. Then for any sequence $D_0, D_1, \ldots$ with $\mathbb{E}|D_N| < \infty$,*

$$\mathbb{E}\left(D_N - D_{N+1}\right) = \left(\frac{1}{m} - 1\right)\left(D_0 - \mathbb{E}D_N\right)$$

Proof By definition of the geometric distribution,

$$\begin{aligned}
&\mathbb{E}\left(D_N - D_{N+1}\right) = \sum_{n\geq 0}\left(D_n - D_{n+1}\right)m^n(1-m) \\
&= (1-m)\left(D_0 - \sum_{n\geq 1} D_n\left(m^{n-1} - m^n\right)\right) = (1-m)\left(\frac{1}{m}D_0 - \sum_{n\geq 0} D_n\left(m^{n-1} - m^n\right)\right) \\
&= (1-m)\left(\frac{1}{m}D_0 - \frac{1}{m}\sum_{n\geq 0} D_n m^n(1-m)\right) = \left(\frac{1}{m} - 1\right)\left(D_0 - \mathbb{E}D_N\right),
\end{aligned}$$

where in the last equality we have used the fact that $\mathbb{E}|D_N| < \infty$. □

To show how geometrization exactly works, we recall that in the proof of 1, we obtained the following bound on the distance $\|x_{k+1} - x^*\|$:

$$\mathbb{E}\left\|x_{k+1} - x^*\right\|^2 \leq \left\|x_k - x^*\right\|^2 - 2\eta(1-2L\eta)\left(f\left(x_k\right) - f^*\right) - \mu\eta\left\|x_k - x^*\right\|^2 + 4L\eta^2\left(f\left(\tilde{x}_s\right) - f^*\right)$$

We then performed a telescoping sum, which we were able to express in terms of an expectation with respect to $N \sim \text{Unif}\{0, \ldots, m-1\}$, to arrive at

$$\mathbb{E}\left\|x_m - x^*\right\|^2 \leq \mathbb{E}\left\|\tilde{x}_s - x^*\right\|^2 - 2\eta(1-2L\eta)m\left(\mathbb{E}f\left(\tilde{x}_{s+1}\right) - f^*\right) - \mu\eta m\mathbb{E}\left\|\tilde{x}_{s+1} - x^*\right\|^2 + 4L\eta^2 m\left(\mathbb{E}f\left(\tilde{x}_s\right) - f^*\right)$$

We notice that the term $\mathbb{E}\left\|x_m - x^*\right\|^2$ appears in the above expression, and unfortunately we cannot relate it to $\tilde{x}_{s+1}$ since N is randomly distributed between 0 and $m-1$. We therefore chose to simply drop the term by lower bounding it by 0 in deriving the result of 1. As a consequence, the optimal choice of m ends up requiring knowledge of μ. However, if we instead take $N \sim \text{Geom}(m)$, we can entirely circumvent this issue. We specifically consider the choice of $m = n/(n+1)$, which we note is entirely independent of μ. Letting $k = N$ and defining $D_N = \mathbb{E}\left\|x_N - x^*\right\|^2$, we apply 2 to obtain

$$\mathbb{E}\left\|\tilde{x}_{s+1} - x^*\right\|^2 \leq \mathbb{E}\left\|\tilde{x}_s - x^*\right\|^2 - 2\eta(1-2L\eta)n\left(\mathbb{E}f\left(\tilde{x}_{s+1}\right) - f^*\right) - \mu\eta n\mathbb{E}\left\|\tilde{x}_{s+1} - x^*\right\|^2 + 4L\eta^2 n\left(\mathbb{E}f\left(\tilde{x}_s\right) - f^*\right).$$

We see that we now conveniently have $\mathbb{E}\left\|\tilde{x}_{s+1} - x^*\right\|^2$ in place of $\mathbb{E}\left\|x_m - x^*\right\|^2$, which allows us to derive a tighter bound. Rearranging yields

$$2\eta(1-2\eta L)\left(\mathbb{E}f\left(\bar{x}_{s+1}\right) - f^*\right) + \left(\frac{1}{n} + \mu\eta\right)\mathbb{E}\left\|\tilde{x}_{s+1} - x^*\right\|^2 \leq 4L\eta^2\left(\mathbb{E}f\left(\tilde{x}_s\right) - f^*\right) + \frac{1}{n}\mathbb{E}\left\|\bar{x}_s - x^*\right\|^2$$

It can then be shown that this result gives the same query complexity as regular SVRG, though we do not include the proof here for sake of brevity.

3 Randomized Coordinate Descent

3.1 Introduction of Randomized Coordinate Descent

We now introduce another stochastic optimization method known as Randomized Coordinate Descent (RCD). We consider the case where our objective function is no longer composite, i.e., it cannot be expressed as $f(x) = \sum_{i=1}^n f_i(x)$. However, we recognize that the gradient can still be decomposed in the following manner by definition:

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n}\right),$$

where n now denotes the dimension of the problem. The key idea of RCD is to only compute a single coordinate of the gradient $\partial f(x)/\partial x_i$, which requires less computational cost than computing the entire gradient $\nabla f(x)$, though we note it may not necessarily be the case that we always obtain a factor of n speedup. The update rule is specifically given by

$$x_{t+1} = x_t - \eta \nabla_{i_t} f(x) e_{i_t},$$

where $i_t \sim \text{Unif}\{1, \ldots, n\}$, $\nabla_{i_t} f(x) = \partial f(x)/\partial x_i$, and e_{i_t} is the standard basis vector along coordinate i_t. We can construct an unbiased estimate of the gradient by appropriate scaling by a factor of n,

$$\tilde{g}(x) = n \nabla_{i_t} f(x) e_{i_t}$$

such that

$$\mathbb{E}\tilde{g}(x) = \sum_{i=1}^{n} \nabla_{i_t} f(x) e_{i_t} = \nabla f(x).$$

We can also characterize the variance,

$$\mathbb{E}\|\tilde{g}(x)\|^2 = \sum_{i=1}^{n} n(\nabla_{i_t} f(x))^2 = n\|\nabla f(x)\|^2,$$

which scales with the dimension of the problem. We recall from our analysis of SGD that if f is Lipschitz, we can bound the variance of the gradient estimator as $\mathbb{E}\|\tilde{g}(x)\|^2 \leq M^2$. We then derived a standard performance guarantee of the form

$$\mathbb{E}f(\bar{x}) - f^* \leq \frac{\|x_0 - x^*\| M}{\sqrt{T}},$$

where $\bar{x} = \frac{1}{T}\sum_{s=1}^{T} x_s$. If we specifically assume that f is 1-Lipschitz, i.e. $\|\nabla f(x)\| \leq 1$, then $\mathbb{E}\|\tilde{g}(x)\|^2 \leq n$, so that $M^2 = n$. Compared to GD, where we simply have $M^2 = 1$, we recognize that RCD is worse by a factor of n, but this is naturally offset by the fact that we only compute one out of the total n coordinates of the gradient at each iteration such that the overall query complexity is essentially the same in this case.

3.2 Randomized Coordinate Descent for Smooth Optimization

In this section, we introduce RCD(γ) algorithm for coordinate-wise smooth functions. First, we formally define the coordinate-wise smoothness.

Assumption 3 (Coordinate-wise Smoothness)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is* coordinate-wise smooth *if there exists $(\beta_1, \beta_2, \ldots, \beta_n)$ s.t.*

$$|\nabla_i f(x + ue_i) - \nabla_i f(x)| \leq \beta_i |u|, \quad \forall i \in [n],\ \forall x \in \mathbb{R}^n,\ \forall u \in \mathbb{R}.$$

When f is twice-differentiable, this is equivalent to

$$(\nabla^2 f(x))_{ii} \leq \beta_i, \quad \forall x \in \mathbb{R}^n.$$

Note that when f is convex, we have

$$\lambda_{\max}(\nabla^2 f(x)) \leq \text{Tr}(\nabla^2 f(x)) \leq \sum_{i=1}^{n} \beta_i,$$

which implies that f is a smooth function with $\beta = \sum_{i=1}^{n} \beta_i$.

With the above coordinate-wise smoothness assumption, we start to analyze the convergence rate of RCD. Recall that GD achieves $O(\frac{1}{T})$ convergence rate for smooth functions with learning rate $\eta = O(\frac{1}{\beta})$.

The choice of η is "conservative" in GD, and thus in RCD we use an "aggressive" algorithm RCD(γ). The updating rule of RCD(γ) is

$$x_{t+1} \leftarrow x_t - \frac{1}{\beta_{i_t}}(\nabla_{i_t} f(x))e_{i_t}$$

where the sampling rule is

$$i_t \sim \mathbb{P}(i_t = i) = P_\gamma(i) = \frac{\beta_i^\gamma}{\sum_{j=1}^n \beta_j^\gamma}.$$

Note that the learning rate in RCD is $\frac{1}{\beta_{i_t}}$ instead of $\frac{1}{\beta} \approx \frac{1}{\sum_{i=1}^n \beta_i}$ as in GD. Therefore, intuitively, the step size of RCD moving in one direction is n times that of GD. Also, to sample i_t, we only need access to a random oracle which returns a uniformly random number between $[0, 1]$, then a binary search with $O(\log n)$ computational time suffices to obtain i_t.

For notation convenience, we also define

$$\|\cdot\|_{[\gamma]} = \sqrt{\sum_{i=1}^n \beta_i^\gamma x_i^2}, \quad \|\cdot\|_{[\gamma]}^* = \sqrt{\sum_{i=1}^n \frac{1}{\beta_i^\gamma} x_i^2}.$$

Now we provide the theoretical guarantee of RCD(γ).

Theorem 4. *For RCD(γ) with $T \geq 2$ and function f that is convex and coordinate-wise smooth (Assumption 3), it holds that*

$$\mathbb{E}[f(x_T)] - f^* \leq \frac{2R_{1-\gamma}^2(x_1)\sum_{i=1}^n \beta_i^\gamma}{T-1},$$

where $R_{1-\gamma}(x_1) = \sup_{x: f(x) \leq f(x_1)} \|x - x^*\|_{[1-\gamma]}$.

Remark The term $R_{1-\gamma}(x_1)$ can be interpreted as the radius. The term $\sum_{i=1}^n \beta_i^\gamma \approx \beta$ when $\gamma \approx 1$. Therefore, the convergence rate of RCD and GD are of the same order. Moreover, the advantage of RCD is that the computational cost per iteration ($O(\log n)$ per iteration with initial cost $O(n)$) is much less than GD ($O(n)$ per iteration).

To prove Theorem 4, we also need the following lemma.

Lemma 5. *For any function f that is convex and coordinate-wise smooth (Assumption 3) and for any i,*

$$\mathbb{E}_{i \sim P_\gamma}[f(x_{t+1}) - f(x_t)] \leq -\frac{1}{2\sum_{i=1}^n \beta_i^\gamma}\left(\|\nabla f(x_t)\|_{[1-\gamma]}^*\right)^2.$$

Proof The proof starts with the following fact

$$f\left(x - \frac{1}{\beta_i}(\nabla_i f(x))e_i\right) - f(x) \leq \frac{-1}{2\beta_i}(\nabla_i f(x))^2, \tag{1}$$

which is due to the coordinate-wise smoothness of f. By choosing $x = x_t$ and taking expectation over both sides, we can obtain that

$$\begin{aligned}
\mathbb{E}_{i \sim P_\gamma}[f(x_{t+1}) - f(x_t)] \leq & \mathbb{E}_{i \sim P_\gamma}\left[\frac{-1}{2\beta_i}(\nabla_i f(x_t))^2\right] \\
= & -\sum_{i=1}^n \left[\frac{P_\gamma(i)}{2\beta_i}(\nabla_i f(x_t))^2\right] \\
= & -\frac{1}{2\sum_{j=1}^n \beta_j^\gamma}\sum_{i=1}^n \left[\frac{1}{\beta_i^{1-\gamma}}(\nabla_i f(x_t))^2\right] \\
= & -\frac{1}{2\sum_{i=1}^n \beta_i^\gamma}\left(\|\nabla f(x_t)\|_{[1-\gamma]}^*\right)^2.
\end{aligned}$$

□

Proof [Proof of Theorem 4] Let $\Delta_t = f(x_t) - f(x^*)$, then by the convexity of f and Holder's inequality, we can obtain that

$$\begin{aligned}\Delta_t \leq & \langle \nabla f(x_t), x_t - x^* \rangle \\ \leq & \|\nabla f(x_t)\|^*_{[1-\gamma]} \|x_t - x^*\|_{[1-\gamma]}.\end{aligned}$$

By (1), we have $f(x_{t+1}) \leq f(x_t) \leq \cdots \leq f(x_1)$, which implies that $\|x_t - x^*\|_{[1-\gamma]} \leq R_{1-\gamma}$. Therefore, by Lemma 5,

$$\begin{aligned}&\mathbb{E}[\Delta_{t+1} - \Delta_t] \leq -\frac{1}{2\sum_{i=1}^n \beta_i^\gamma} \left[\frac{\Delta_t^2}{(R_{1-\gamma}(x_1))^2}\right] \\ \Longrightarrow &\mathbb{E}[\frac{1}{\Delta_t} - \frac{1}{\Delta_{t+1}}] \leq -\frac{1}{2(R_{1-\gamma}(x_1))^2 \sum_{i=1}^n \beta_i^\gamma} \frac{\Delta_t}{\Delta_{t+1}} \leq -\frac{1}{2(R_{1-\gamma}(x_1))^2 \sum_{i=1}^n \beta_i^\gamma}.\end{aligned}$$

Finally, by telescoping, we can conclude that

$$\mathbb{E}[\Delta_t] \leq \frac{2R_{1-\gamma}^2(x_1) \sum_{i=1}^n \beta_i^\gamma}{T-1}.$$

□

3.3 Randomized Coordinate Descent for Smooth and Strongly Convex Optimization

For smooth and strognly convex function, GD needs $O(\kappa \log(1/\varepsilon))$ iterations to achieve ε-accuracy. The following theorem shows that the RCD achieves the similar convergence rate.

Theorem 6. *For function f that is μ-strongly convex w.r.t. $\|\cdot\|_{[1-\gamma]}$ and coordinate-wise smooth (Assumption 3), define $\kappa_\gamma = \frac{\sum_{i=1}^n \beta_i^\gamma}{\mu}$, then RCD($\gamma$) guarantees that*

$$\mathbb{E}[f(x_{t+1}) - f(x^*)] \leq \left(1 - \frac{1}{\kappa_\gamma}\right)^t (f(x_1) - f(x^*)).$$

Remark When $\gamma = 1$ and $L = \sum_{i=1}^n \beta_i$, we have $\kappa_\gamma = \frac{L}{\mu}$, which is consistent with GD.

To prove Theorem 6, we first present the following lemma.

Lemma 7. *Let f be μ-strongly convex w.r.t. any norm $\|\cdot\|$, then*

$$f(x) - f(x^*) \leq \frac{1}{2\mu} \|\nabla f(x)\|_*^2.$$

Proof For any x, y, by the strong convexity and Holder's inequality, we have

$$\begin{aligned}f(x) - f(y) \leq & \nabla f(x)^T (x - y) - \frac{\mu}{2}\|x - y\|^2 \\ \leq & \|\nabla f(x)\|_* \|x - y\| - \frac{\mu}{2}\|x - y\|^2.\end{aligned}$$

Denote $z = \|x - y\|$, then

$$f(x) - f(y) \leq z\|\nabla f(x)\|_* - \frac{\mu}{2}z^2 \leq \frac{1}{2\mu}\|\nabla f(x)\|_*^2.$$

Choosing $y = x^*$ completes the proof. □

Proof [Proof of Theorem 6] Recall Lemma 5 that for any i, we have

$$\mathbb{E}[f(x_{t+1}) - f(x_t)] \leq -\frac{1}{2\sum_{i=1}^n \beta_i^\gamma}\left(\|\nabla f(x_t)\|_{[1-\gamma]}^*\right)^2.$$

By Lemma 7, we can further obtain that

$$\begin{aligned}
&\mathbb{E}[f(x_{t+1}) - f(x_t)] \leq -\frac{\mu}{\sum_{i=1}^n \beta_i^\gamma}(f(x_t) - f(x^*)) \\
\Longrightarrow &\mathbb{E}[\Delta_{t+1} - \Delta_t] \leq -\frac{1}{\kappa_\gamma}\Delta_t \\
\Longrightarrow &\mathbb{E}[\Delta_{t+1}] \leq (1 - \frac{1}{\kappa_\gamma})\mathbb{E}[\Delta_t],
\end{aligned}$$

which concludes that

$$\mathbb{E}[f(x_{t+1}) - f(x^*)] \leq \left(1 - \frac{1}{\kappa_\gamma}\right)^t (f(x_1) - f(x^*)).$$

□

3.4 Summary

In this section, we introduced the RCD(γ) algorithm for coordinate-wise smooth and convex functions, which achieves the same convergence speed as GD while the computational cost per iteration is only $O(\log n)$ compared to $O(n)$ per iteration for GD.

For coordinate-wise smooth and strongly convex functions, RCD(γ) algorithm also achieves the same convergence speed as GD while the computational cost per iteration is exponentially better than GD.

References

[1] L. Lei and M. I. Jordan, "On the adaptivity of stochastic gradient-based optimization," *SIAM Journal on Optimization*, vol. 30, no. 2, pp. 1473–1500, jan 2020. [Online]. Available: https://doi.org/10.1137%2F19m1256919