ECE 901: Large-scale Machine Learning and Optimization **Fall 2016**

Lecture 7 — 09/27

Lecturer: Dimitris Papailiopoulos *Scribe: Alisha Zachariah*

7.1 Re-examining Stochastic Gradient Descent

The motivating idea behind *Stochastic Gradient Descent* (SGD) is ensuring cheap per iteration complexity.

Recall the set-up:

$$\min \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

$$\text{Update: } x_{k+1} = x_k - \gamma \nabla f_{s_k}(x_k)$$

While the cost per iteration of SGD is $\approx n$ times faster than *Gradient Descent* (GD), the iteration complexity of SGD is significantly worse (exponentially worse, see 6.1.4).

The goal of faster iteration complexity while maintaining efficiency per iteration motivates *Stochastic Variance Reducing Gradient* (SVRG). The derivation of convergence bounds for SGD motivates the design of the SVRG algorithm.

7.1.1 Convergence bounds for SGD:

The convergence rates we derived previously suggest a worst case convergence rate of $\sim 1/T$ for SGD, where T is the number of iterations. In practice, the actual convergence rate may be somewhat better than this bound. Infact, we should expect SGD to begin with linear convergence comparable to GD, eventually slowing down to the $1/T$ rate predicted by our analysis (Fig. 7.1). We will develop the intution for why to expect this. SVRG is intended to address this slowing down.

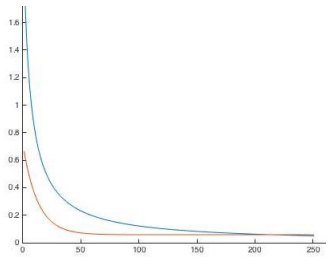

Figure 7.1. Convergence rates: Worst-case bound in blue vs. true rate in red.

We will make the following assumptions regarding he objective function f:

- λ-strong convexity: $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \lambda \|x - y\|^2$
- β-smoothness: $\|\nabla f(x) - \nabla f(y)\|^2 \leq \beta \|x - y\|^2$
- $E_s \|\nabla f_s(x)\| \leq M^2$

Notation : $\Delta_k = \|x_k - x^*\|^2$
Recall that using λ-strong convexity we get

$$\mathbb{E}\Delta_{k+1} \leq \underbrace{(1 - \gamma\lambda)\mathbb{E}\Delta_k}_{\text{const. factor decrease}} + \underbrace{\gamma^2 \mathbb{E}\|\nabla f_{s_k}(x_k)\|^2}_{\text{variance of gradient, causes slowing}}$$

Also recall, at this point in the analysis of GD we use β-smoothness and that $\nabla f(x^*) = 0$ to get constant factor decrease:

$$\begin{aligned}
\Delta_{k+1} &\leq (1 - \gamma\lambda)\Delta_k + \gamma^2 \|\nabla f(x_k)\|^2 \\
&= (1 - \gamma\lambda)\Delta_k + \gamma^2 \|\nabla f(x_k) - \nabla f(x^*)\|^2 \quad (\nabla f(x^*) = 0) \\
&\leq (1 - \gamma\lambda)\Delta_k + \gamma^2 \beta^2 \Delta_k \quad (\beta\text{-smoothness}) \\
&= (1 - \gamma\lambda + \gamma^2\beta^2)\Delta_k
\end{aligned}$$

Replicating this idea for SGD:

$$\begin{aligned}
\mathbb{E}\Delta_{k+1} &\leq (1 - \gamma\lambda)\mathbb{E}\Delta_k + \gamma^2 \mathbb{E}\|\nabla f_{s_k}(x_k)\|^2 \\
&= (1 - \gamma\lambda)\mathbb{E}\Delta_k + \gamma^2 \mathbb{E}\|\nabla f_{s_k}(x_k) - \nabla f_{s_k}(x^*) + \nabla f_{s_k}(x^*)\|^2 \\
&= (1 - \gamma\lambda)\mathbb{E}\Delta_k + 2\gamma^2 \mathbb{E}\|\nabla f_{s_k}(x_k) - \nabla f_{s_k}(x^*)\|^2 + 2\gamma^2 \mathbb{E}\|\nabla f_{s_k}(x^*)\|^2 \text{ (by 7.2)}
\end{aligned}$$

where the β-smooth assumption

$$\mathbb{E}\Delta_{k+1} \leq \underbrace{(1 - \gamma\lambda + 2\beta^2\gamma^2)\mathbb{E}\Delta_k}_{A} + \underbrace{2\gamma^2 \mathbb{E}\|\nabla f_{s_k}(x^*)\|^2}_{B} \tag{7.1}$$

Initially $B << A$ and we observe the linear rate regime, once $B > A$ we observe $1/T$-rate.

7.2 Stochastic Variance Reducing Gradient

The idea behind *Stochastic Variance Reducing Gradient* (SVRG) is to ensure decaying "variance" and SGD-like cost/iteration.

$$\text{Update: } v_k$$
$$x_{k+1} = x_k - \gamma v_k$$

We would like to pick v_k such that $\mathbb{E}v_k = \nabla f$, ie. we would like unbiased "gradients".

$$v_k = \nabla f_{s_k}(x_k) - \nabla f_{s_k}(y) + \underbrace{\nabla f(y)}_{\text{ensures unbiased}}$$

The terms in red should be variance reducing. For now y is fixed to minimize the number of full gradient computations and needs to be chosen.

7.2.1 Analysis of SVRG

Let's analyze the "variance" :

$$\begin{aligned}
\mathbb{E}_s\|v_k\|^2 &= \mathbb{E}\|\nabla f_s(x) - \nabla f_s(y) + \nabla f(y)\|^2 \\
&= \mathbb{E}\|\nabla f_s(x) - \nabla f_s(y) + \nabla f(y) \pm \nabla f_s(x^*)\|^2 \\
&\leq 2\mathbb{E}\|\nabla f_s(x) - \nabla f_s(x^*)\|^2 + 2\mathbb{E}\|\nabla f_s(y) - \nabla f_s(x^*) - \nabla f(y)\|^2 \text{ (by 7.2)} \\
&\leq 2\beta^2\mathbb{E}\Delta_k + 2\mathbb{E}\|\nabla f_s(y) - f_s(x^*)\|^2 \text{ (by 7.3)} \\
&\leq 2\beta^2\mathbb{E}\Delta_k + 2\beta^2\mathbb{E}\|y - x^*\|^2 \quad (\beta - \text{smoothness})
\end{aligned}$$

We have effectively introduced y into the bound for variance. There are now two competing forces at play: picking a fresh y more often should decrease the variance, however doing this too often involves computing too many full gradients. Let's set $y = x_1$ and see what happens...

Substituting back into 7.1:

$$\begin{aligned}
\mathbb{E}\Delta_{k+1} &\leq \mathbb{E}\Delta_k - \gamma\lambda\mathbb{E}\Delta_k + 2\gamma^2\beta^2\mathbb{E}\Delta_k + 2\beta^2\mathbb{E}\Delta_1 \\
&= (1 - \gamma\lambda + 2\gamma^2\beta^2)\mathbb{E}\Delta_k + 2\gamma^2\beta^2\mathbb{E}\Delta_1
\end{aligned}$$

Unrolling this:

$$\begin{aligned}
\mathbb{E}\Delta_{k+1} &\leq (1 - \gamma\lambda + 2\gamma^2\beta^2)\mathbb{E}\Delta_k + 2\gamma^2\beta^2\mathbb{E}\Delta_1 \\
&\leq (1 - \gamma\lambda + 2\gamma^2\beta^2)^2\mathbb{E}\Delta_{k-1} + \underbrace{(1 - \gamma\lambda + 2\gamma^2\beta^2)}_{*} 2\gamma^2\beta^2\mathbb{E}\Delta_1 + 2\gamma^2\beta^2\mathbb{E}\Delta_1 \\
&\leq (1 - \gamma\lambda + 2\gamma^2\beta^2)^2\mathbb{E}\Delta_{k-1} + 2\gamma^2\beta^2\mathbb{E}\Delta_1 + 2\gamma^2\beta^2\mathbb{E}\Delta_1 \quad (* \text{ must be} < 1) \\
&\vdots \\
&\leq (1 - \gamma\lambda + 2\gamma^2\beta^2)^k\mathbb{E}\Delta_1 + 2k\gamma^2\beta^2\mathbb{E}\Delta_1
\end{aligned}$$

Suppose we would like this to be $\leq 0.5\mathbb{E}\Delta_1$ after T iterations of SVRG, what is the best choice of T and γ?

If we pick $\gamma = O(1)\lambda/\beta^2$, then it turns out that we can set $T = O(1)\beta^2/\lambda^2$. This can be improved to $T = O(1)\beta/\lambda$. Note that β/λ is the condition number κ.

This is one ***epoch*** of the algorithm. If E is the total number of epochs executed we have

$$\mathbb{E}\Delta_k \leq (0.5)^E \,.\, \mathbb{E}\Delta_1$$

7.2.2 Algorithmic complexity

Algorithm 1 SVRG

$y \leftarrow x_0$
$k \leftarrow 0$
for e = 1:E **do**
  $g \leftarrow \nabla f(y)$ *(full gradient)*
  for s = 1:S **do**
    $x_k \leftarrow x_k - \gamma(\nabla f_s(x_k) - \nabla f_s(y) + g)$ *(smaller gradient eval)*
    $k \leftarrow k + 1$
  end for
  $y \leftarrow x_{k-1}$
end for

- **SVRG:** $E \sim \log(\frac{1}{\epsilon})$ so the complexity is $O((n+\kappa)\log(\frac{1}{\epsilon}))$
- **GD:** $T \sim \kappa \log(\frac{1}{\epsilon})$ so the complexity is $O(n\kappa \log(\frac{1}{\epsilon}))$
- **SGD:** $T \sim \frac{\kappa}{\epsilon}$ and the complexity is $O(\frac{\kappa}{\epsilon})$

The moral here is that even though we are allowing ourselves a few gradient computations here, we don't really pay too much in terms of complexity.

The big issue in SVRG is the extra parameter, namely the number of epochs. Another issue it parallelizability – SVRG cannot be parametrized so easily since the updates contain some history.

7.2.3 Concentration

Our analysis demonstrates that we can make $\mathbb{E}\Delta_k$ small. We can convert this into a statement about the probability that Δ_k is small using Markov's inequality:

$$\text{Markov's inequality: If } X \geq 0, \;\; \mathbb{P}(X \geq a\mathbb{E}(X)) \leq \frac{1}{a}$$

$$\text{So} \quad \mathbb{P}(\Delta_k \geq 10\mathbb{E}\Delta_k) \leq \frac{1}{10}$$

So if we run SVRG r times, then the probability that Δ_k is large in all r trials is less that 10^{-r} and we can get concentration in this way.

However, can we get exponential concentration, ie. $\mathbb{P}(\Delta_k \geq a\mathbb{E}\Delta_k) \leq e^{-a}$? It may be possible to achieve this by proving some version of Hoeffding's inequality for strongly convex loss functions.

7.3 Conclusion

There are several versions of SVRG (SAG/SAGA, SDCA, Finito) with variations and applications to different kinds of problems (e.g. faster convergence to saddle points in non-convex cases). There is a lot of interest among the NIPS/ICML communities and lots of active research on how to improve SVRG.

7.4 Appendix:

- $$\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2 \tag{7.2}$$

  This follows from the Parallelogram Law $\|a+b\|^2 + \|a-b\|^2 = 2\|a\|^2 + 2\|b\|^2$, which is easily verified.

- $$\mathbb{E}\|Y - \mathbb{E}Y\|^2 \leq \mathbb{E}\|Y\|^2 \tag{7.3}$$

$$\begin{aligned}
\mathbb{E}\|Y - \mathbb{E}(Y)\|^2 &= \mathbb{E}\sum_i (Y_i - \mathbb{E}(Y_i))^2 \\
&= \sum_i \mathbb{E}(Y_i - \mathbb{E}(Y_i))^2 = \sum_i \mathbb{E}(Y_i^2) - \mathbb{E}(Y_i)^2 \\
&\leq \sum_i \mathbb{E}(Y_i^2) = \mathbb{E}\sum_i Y_i^2 \\
&= \mathbb{E}\|Y\|^2
\end{aligned}$$