Lecture 15

Scribe?

Last time

- Stochastic Gradient Descent
- Examples
- Analysis

Today

- Analysis continued
- Convex guarantees
- Extensions

Theorem Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and $g(x, z)$ is an unbiased estimator such that

$$\mathbb{E}\left[\| g(x,z) - \nabla f(x)\|^2\right] \leq \sigma^2 \quad \forall x.$$

Then the iterates of stochastic gradient descent with $0 < \alpha_k < 2/L$ satisfy

$$\mathbb{E}\left[\min_{k \leq T} \|\nabla f(x_i)\|_2^2\right] \leq \frac{(f(x_0) - \min f) + \frac{\sigma^2 L}{2}\sum_{k=0}^{T} \alpha_k^2}{\sum_{k=0}^{T} \alpha_k\left(1 - \frac{L\alpha_k}{2}\right)}$$

Relevant properties of the expectation

- Linearity

Given $X_1, \dots, X_n$ r.v. and constants $\lambda_1, \dots, \lambda_n$, we have

$$\mathbb{E}\left[\sum_{i=1}^{n} \lambda_i X_i\right] = \sum \lambda_i \mathbb{E} X_i.$$

▷ Tower law

Given two random variables X, Y

$$\mathbb{E}_X\left[\mathbb{E}[Y \mid X]\right] = \mathbb{E}[Y]$$

conditional expectation (↑ $\mathbb{E}[Y \mid X]$)

Proof: By the Taylor Approximation Theorem

$$f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2} \|x_{k+1} - x_k\|^2$$
$$= f(x_k) - \alpha_k \nabla f(x_k)^T g_k + \frac{L\alpha_k^2}{2} \|g_k\|^2$$

Conditioning on x_k

random because of z_k (↓ g_k)

$$\mathbb{E}[f(x_{k+1}) \mid x_k] \leq f(x_k) - \alpha_k \mathbb{E}\left[\nabla f(x_k)^T g_k \mid x_k\right] + \frac{L\alpha_k^2}{2} \mathbb{E}\left[\|g_k\|^2 \mid x_k\right]$$

Linearity ↓

$$= f(x_k) - \alpha_k \nabla f(x_k)^T \mathbb{E}[g_k \mid x_k] + \frac{L\alpha_k^2}{2} \mathbb{E}\left[\|g_k\|^2 \mid x_k\right]$$

$$(\heartsuit) \quad \le f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \frac{L\alpha_k^2}{2}\left[\sigma^2 + \|\nabla f(x_k)\|^2\right]$$

$$= f(x_k) - \left(\alpha_k + \frac{L\alpha_k^2}{2}\right)\|\nabla f(x_k)\|^2 + \frac{L\alpha_k^2}{2}\sigma^2.$$

By Tower Law

$$\mathbb{E}\left[f(x_{k+1})\right] \le \mathbb{E} f(x_k) - \left(\alpha_k + \frac{L\alpha_k^2}{2}\right)\mathbb{E}\|\nabla f(x_k)\|^2 + \frac{L\alpha^2_k}{2}\sigma^2.$$

By recursively applying this formula

$$\mathbb{E}\left[f(x_{T+1})\right] \le \mathbb{E} f(x_0) - \sum_{k=0}^{T}\left(\alpha_k - \frac{L\alpha_k^2}{2}\right)\mathbb{E}\|\nabla f(x_k)\|^2 + \sum_{k=0}^{T}\frac{L\alpha_k^2\sigma^2}{2}$$

The result follows from reordeng and using the fact that

$$\mathbb{E}\left[\min_{k\le T}\|\nabla f(x_k)\|^2\right]\sum_{k=0}^{T}\left(\alpha_k - \frac{L\alpha_k^2}{2}\right)$$

$$\le \sum_{k=0}^{T}\left(\alpha_k - \frac{L\alpha_k^2}{2}\right)\mathbb{E}\left[\|\nabla f(x_k)\|^2\right]. \quad \square$$

Consequences

If $\alpha_k = \frac{1}{L\sqrt{T+1}} \Rightarrow 1 - \frac{L\alpha_k}{2} \geq \frac{1}{2}$.

Thus we derive

$$\mathbb{E}\left[\min_{k\leq T} \|\nabla f(x_k)\|^2\right] \leq \frac{(f(x_0) - \min f) + \frac{\sigma^2}{2L}}{\frac{1}{2}\sqrt{T+1}}$$

$$= O\left(\frac{1}{\sqrt{T}}\right).$$

By Jensen's inequality

$$\Rightarrow \mathbb{E} \min_{k\leq T} \|\nabla f(x_k)\| = O(T^{-1/4}).$$

This is rather slow, however it improves when have convexity.

Convex guarantees

<u>Theorem</u> Consider the same setting as the previous Theorem, further assume that $\alpha_k = \alpha \leq \frac{1}{L}$, f is convex and $x^* \in \operatorname{argmin} f$. Then

$$\mathbb{E}\left[\min_{k\leq T}\{f(x_k) - f(x^*)\}\right] \leq \frac{\|x_0 - x^*\|^2}{2\alpha(k+1)} + \alpha\sigma_*^2$$

In particular if $\alpha = \frac{1}{\sqrt{T+1}}$ and $T \geq L^2$

$$\mathbb{E}\left[\min_{k \leq T} \{ f(x_k) - f(x^*) \}\right] \leq \frac{\|x_0 - x^*\|^2 + 2\sigma^2}{2\sqrt{k+1}}.$$

$\top$

Proof When $\alpha \leq \frac{1}{L}$, ($\heartsuit$) gives

$$\mathbb{E}[f(x_{k+1}) | x_k] \leq f(x_k) - \frac{\alpha}{2} \|\nabla f(x_k)\|^2 + \frac{\alpha\sigma^2}{2}$$

By convexity

By Assumption

$\mathbb{E}[\|g(x,z)\|^2 | x] - \sigma^2 \leq \|\nabla f(x)\|^2$

$$\leq f(x^*) - \nabla f(x_k)^T (x^* - x_k) - \frac{\alpha}{2} \mathbb{E}[\|g(x_k, z)\|^2 | x_k] + \alpha\sigma^2$$

$$\leq f(x^*) - \mathbb{E}\left[g(x_k, z)^T (x^* - x_k) + \frac{\alpha}{2} \|g(x_k, z)\|^2 | x_k\right] + \alpha\sigma^2$$

Using that

$\|x_{k+1} - x^*\|^2 = \|x_k - x^* - \alpha g_k\|^2 = \|x - x^*\|^2 + 2\alpha g_k^T (x^* - x_k) + \alpha^2 \|g_k\|^2$

$$\leq f(x^*) - \mathbb{E}\left[\frac{1}{2\alpha}\left[\|x_{k+1} - x^*\|^2 - \|x_k - x^*\|^2\right] | x_k\right] + \alpha\sigma^2$$

By Tower law

$$\mathbb{E}[f(x_{k+1}) - f(x^*)] \leq \frac{1}{2\alpha} \mathbb{E}[\|x_{k+1} - x^*\|^2 - \|x_k - x^*\|^2] + \alpha\sigma^2.$$

Once more the result follows by summing up and dividing by T. $\square$

Remark

- The rate above is of the order $O(\frac{1}{\sqrt{T}})$, exactly like the rate for nonsmooth convex optimization.
- In HW 4 you'll show the same rate for stochastic nonsmooth convex opt. There, we will have $g(x, z)$ s.t. $\mathbb{E}[g(x, z)] \in \partial f(x)$.

Extensions

Acceleration?

The noise dominates and leads to slow convergence. Best known rate

$$O\left(\frac{L\|x_0 - x^*\|^2}{T^2} + \frac{\sigma^2}{\sqrt{T}}\right).$$

Randomized coordinate descent

Assume our oracle is

(:) $$i \sim \mathrm{Unif}(\{1, \dots, d\})$$
$$g(x,i) = d \cdot \frac{\partial f}{\partial x_i}(x) \cdot e_i .$$

The analysis above yields a guarantee but we can do better.

Theorem. Assume $f: \mathbb{R}^d \to \mathbb{R}$ L-smooth. Then SGD with (:) and $\alpha_k = \frac{1}{Ld}$ yields

$$\mathbb{E}\left[\min_{k \le T} \|\nabla f(x_k)\|^2\right] \le \frac{2Ld(f(x_0) - \min f)}{T}.$$

Proof Indeed this oracle gives descent

At iter k,

$$f(x_{k+1}) \le f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

$$= f(x_k) - \frac{1}{Ld} d \frac{\partial f}{\partial x_i}(x_k) \cdot \nabla f(x_k)^T e_i + \frac{1}{2Ld^2}\left(d \frac{\partial f}{\partial x_i}(x_k)\right)^2$$

$$= f(x) - \frac{1}{2L}\left(\frac{\partial f(x_k)}{\partial x_i}\right)^2.$$

Taking expectations

$$\mathbb{E}[f(x_{k+1})] \leq \mathbb{E}[f(x_k)] - \frac{1}{2L}\mathbb{E}\left[\left(\frac{\partial f}{\partial x_i}(x_k)\right)^2\right]$$

$$= \mathbb{E}[f(x_k)] - \frac{1}{2L}\frac{1}{d}\mathbb{E}[\|\nabla f(x_k)\|^2]$$

$$\mathbb{E}\left[\left(\frac{\partial f}{\partial x_i}(x)\right)^2 \Big| x\right] = \frac{1}{d}\|\nabla f(x_k)\|^2.$$

By recursively applying the formula above, we obtain

$$\mathbb{E}[f(x_{T+1})] \leq \mathbb{E}[f(x_0)] - \frac{1}{2Ld}\sum_{k=0}^{T}\mathbb{E}[\|\nabla f(x_k)\|^2].$$

Reordering and multiplying by $\frac{1}{T}$, yields

$$\mathbb{E}\left[\min_{k\leq T}\|\nabla f(x_k)\|^2\right] \leq \frac{2Ld\,(f(x_0) - \min f)}{T}.$$

□

This is the same rate as in the deterministic case.

Extensions to greedy and cyclic rules can be found in [Nutini, ICML'15] and [Beck, Tetruashvili, SIOPT 15'].

Stochastic Variance Reduced Gradient (SVRG)

Recall the finite sum problem

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

The SVRG reads as follows

Algorithm

Set $\tilde{x}_0 \leftarrow x_0$

for $i = 0, \dots$

  $y_0 \leftarrow \tilde{x}$

  for $j = 0, \dots, 2d$

    Draw $\ell \sim \mathrm{Unif}(\{1, \dots, n\})$

    $g_j \leftarrow \nabla f(\tilde{x}_i) + \nabla f_\ell(y_j) - \nabla f_\ell(\tilde{x})$

    $y_{j+1} \leftarrow y_j - \alpha g_j$

  end for

  $\tilde{x}_{i+1} \leftarrow \frac{1}{2d+1}\sum_{j=0}^{2d} y_j$.

end for

(outer loop: the outer for … end for; inner loop: the inner for … end for)

Theorem: Assume $f:\mathbb{R}^d \to \mathbb{R}$ L-smooth μ-strongly convex. Then, if α sufficiently small

$\gamma \in (0,1)$.

$$\mathbb{E}[f(\tilde{x}_k) - \min f] \leq \gamma^k [f(\tilde{x}_0) - \min f].$$

Proof: [Johnson, Zhang 2013] $\square$