Non-Convex Optimization

Zed

June 28, 2017

0.1 Deterministic Version

Consider the problem $\min_{x\in\mathbb{R}^n} f(x)$, where f is not necessarily convex. And we assume the gradient of f is Lipschitz-continuous $\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|$, $\forall x, y$.

In such a formulation we cannot guarantee that the algorithm converges to optimum. So we are interest in, alternatively, the rate in which the gradient goes to zero, i.e. the algorithm converges to a .. point.

By smoothness we have

$$\begin{aligned}
f(x_{t+1}) &\leq f(x_t) + \nabla f(x_t)^\top (x_{t+1} - x_t) + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\
&= f(x_t) - \gamma_t \|\nabla f(x_t)\|^2 + \frac{L\gamma_t^2}{2}\|\nabla f(x_t)\|^2 \\
&= f(x) - \gamma_t\left(1 - \frac{L\gamma_t}{2}\right)\|\nabla f(x_t)\|^2
\end{aligned} \tag{1}$$

$\Rightarrow \gamma_t(1 - \frac{L\gamma_t}{2})\|\nabla f(x_t)\|^2 \leq f(x_t) - f(x_{t+1})$. Taking summation:

$$\sum_{t=1}^{k} \gamma_t(1 - \frac{L\gamma_t}{2})\|\nabla f(x_t)\|^2 \leq f(x_1) - f(x_{k+1}) \leq f(x_1) - f^* \tag{2}$$

Pick output $\overline{x}_k$, such that $\|\nabla f(\overline{x}_k)\| = \min_{t=1,\ldots,k} \|\nabla f(\overline{x}_t)\|$. So

$$\begin{aligned}
\sum_{t=1}^{k} \gamma_t\left(1 - \frac{L\gamma_t}{2}\right)\|\nabla f(x_t)\|^2 &\geq \|\nabla f(\overline{x}_k)\|^2 \sum_{t=1}^{k}\gamma_t\left(1 - \frac{L\gamma_t}{2}\right) \\
\|\nabla f(\overline{x}_k)\|^2 &\leq \frac{f(x_1) - f^*}{\sum_{t=1}^{k}\gamma_t(1 - L\gamma_t/2)}
\end{aligned} \tag{3}$$

If $\gamma_t = 1/L$, then $\|\nabla f(\overline{x}_k)\|^2 \leq \frac{2(f(x_1)-f^*)}{k}$.

0.2 Stochastic Version

$x_{t+1} = x_t - \gamma_t G(x_t, \xi_t)$; $\delta_t = \nabla f(x_t) - G(x_t, \xi_t)$. $\mathbb{E}[\delta_t] = 0$, $\mathbb{E}\left[\|\delta_t\|^2\right] = \sigma^2$.

$$\begin{aligned}
f(x_{t+1}) &\leq f(x_t) + \nabla f(x_t)^\top (x_{t+1} - x_t) + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\
&= f(x_t) - \gamma_t \nabla f(x_t)^\top [\nabla f(x_t) - \delta_t] + \frac{L\gamma_t^2}{2}\|\nabla f(x_t) - \delta_t\|^2 \\
&= f(x) - \gamma_t\|\nabla f(x_t)\|^2 + \gamma_t \nabla f(x_t)^\top \delta_t + \frac{L\gamma_t^2}{2}\left(\|\nabla f(x_t)\|^2 - 2\nabla f(x_t)^\top \delta_t + \|\delta_t\|^2\right)
\end{aligned} \tag{4}$$

Take expectation:

$$\mathbb{E}\left[f(x_{t+1})\right] \leq \mathbb{E}\left[f(x_t)\right] - \gamma_t(1 - L\gamma_t/2)\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + \frac{L\sigma^2\gamma_t^2}{2} \tag{5}$$

Take summation:

$$\begin{aligned}\sum_{t=1}^{k}\gamma_t(1 - L\gamma_t/2)\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] &\leq f(x_1) - \mathbb{E}\left[f(x_{k+1}\right] + \sum_{t=1}^{k}\frac{L\sigma^2\gamma_t^2}{2} \\ &\leq f(x_1) - f^* + \sum_{t=1}^{k}\frac{L\sigma^2\gamma_t^2}{2}\end{aligned} \tag{6}$$

We run the algorithm for k iterations. Randomly pick a solution x_R as the output, such that

$$\mathbb{P}\left(R = t\right) = \frac{\gamma_t(1 - L\gamma_t/2)}{\sum_{t=1}^{k}\gamma_t(1 - L\gamma_t/2)}$$

And then we find that

$$\mathbb{E}\left[\|\nabla f(x_R)\|^2\right] = \frac{\sum_{t=1}^{k}\gamma_t(1 - L\gamma_t/2)\mathbb{E}\left[\|\nabla f(x_t)\|^2\right]}{\sum_{t=1}^{k}\gamma_t(1 - L\gamma_t/2)} \leq \frac{f(x_1) - f^* + \sum_{t=1}^{k}\frac{L\sigma^2\gamma_t^2}{2}}{\sum_{t=1}^{k}\gamma_t(1 - L\gamma_t/2)} \tag{7}$$

Take $\gamma_t = \gamma \leq 1/L$, $t = 1, ..., k$, then

$$\text{RHS of (7)} \leq \frac{f(x_1) - f^* + k\frac{L\sigma^2\gamma^2}{2}}{k\gamma/2} = \frac{2(f(x_1) - f^*)}{\gamma k} + \gamma L\sigma^2 \tag{8}$$

$$\gamma = \min\left\{1/L, \sqrt{\frac{2L[f(x_1) - f^*]}{L^{-2}k}}\right\} \tag{9}$$

$$\mathbb{E}\left[\|\nabla f(x_R)\|^2\right] \leq \frac{2L(f(x_1) - f^*)}{k} + 2\sigma\sqrt{\frac{2L[f(x_1) - f^*]}{L^{-2}k}} \tag{10}$$