Notes

Zed

June 28, 2017

1 Stochastic Problem & Strong Convexity

$\mathbb{E}[G(x_t, \xi_t)] = \nabla f(x_t)$.
Define the error $\delta_t := \nabla f(x_t) - G(x_t, \xi_t)$, and make following assumptions:

1. $\mathbb{E}[\delta_t] = 0$, and δ_t is independent of δ_t.

2. $\mathbb{E}\left[\|\delta_t\|^2\right] = \sigma^2$.

The subproblem for each iteration is

$$x_{t+1} = \mathrm{argmin}_{x \in X} \gamma_t \langle G(x_t, \xi_t), x\rangle + \frac{1}{2}\|x - x_t\|^2$$

And the optimality condition becomes:

$$\gamma_t \langle G(x_t, \xi_t), x_{t+1} - x\rangle \leq \frac{1}{2}\left(\|x - x_t\|^2 - \|x - x_{t+1}\|^2 - \|x_t - x_{t+1}\|^2\right)$$

call it (OPT2').

$$\begin{aligned}
f(x_{t+1}) &\leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\
&= \underbrace{f(x_t) + \langle \nabla f(x_t), x - x_t\rangle}_{\text{convexity}} + \langle \nabla f(x_t), x_{t+1} - x\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\
&\leq f(x) + \langle \underbrace{\nabla f(x_t) - \delta_t}_{G(x_t,\xi_t)} + \delta_t, x_{t+1} - x\rangle + \frac{1}{2}\|x_{t+1} - x_t\|^2 \\
&\leq f(x) + \underbrace{\langle G(x_t, \xi_t), x_{t+1} - x\rangle}_{\text{OPT2'}} + \langle \delta_t, x_{t+1} - x\rangle + \frac{1}{2}\|x_{t+1} - x_t\|^2 \\
&\leq f(x) + \frac{1}{2\gamma_t}\left[\|x - x_t\|^2 - \|x - x_{t+1}\|^2\right] + \langle \delta_t, x_{t+1} - x_t\rangle - \frac{1}{2}\left(\frac{1}{\gamma_t} - L\right)\|x_{t+1} - x_t\|^2
\end{aligned} \tag{1}$$

Note that $\mathbb{E}[\langle \delta_t, x_{t+1} - x\rangle] \neq 0$[1], since δ_t is dependent on $x_{t+1} - x$. However it is independent to $x_t - x$, so we want to replace x_{t+1} with x_t. Consider:

$$\begin{aligned}
&\langle \delta_t, x_{t+1} - x\rangle - \frac{1}{2}\left(\frac{1}{\gamma_t} - L\right)\|x_{t+1} - x_t\|^2 \\
&= \langle \delta_t, x_t - x\rangle + \langle \delta_t, x_{t+1} - x_t\rangle - \frac{1}{2}\left(\frac{1}{\gamma_t} - L\right)\|x_{t+1} - x_t\|^2 \\
&\leq \langle \delta_t, x_t - x\rangle + \frac{\|\delta_t\|^2}{2(\gamma_t^{-1} - L)} \quad (\text{with } \gamma_t < 1/L) \;\; (\dagger)
\end{aligned} \tag{2}$$

[1] $\mathbb{E}[\langle \delta_t, x_{t+1} - x\rangle] \neq \langle \mathbb{E}[\delta_t], \mathbb{E}[x_{t+1} - x]\rangle$

$\Rightarrow$

$$\begin{aligned}\mathbb{E}\left[(\dagger)\right] &= \mathbb{E}\left[\langle \delta_t, x_t - x\rangle + \frac{\|\delta_t\|^2}{2(\gamma_t^{-1} - L)}\right] \\ &= 0 + \mathbb{E}\left[\frac{\|\delta_t\|^2}{2(\gamma_t^{-1} - L)}\right] \le \frac{\sigma^2}{2(\gamma_t^{-1} - L)}\end{aligned} \tag{3}$$

Combine (1) and (3), we have, for $t = 1, 2, ..., k$:

$$\gamma_t \mathbb{E}\left[f(x_{t+1}) - f(x)\right] \le \frac{1}{2}\left[\|x - x_t\|^2 - \|x - x_{t+1}\|^2\right] + \frac{\gamma_t^2\sigma^2}{2(1 - L\gamma_t)} \tag{4}$$

Take summation (telescoping) of the equation above:

$$\sum_{t=1}^{k} \gamma_t \mathbb{E}\left[f(x_{t+1}) - f(x)\right] \le \frac{1}{2}\|x - x_1\|^2 + \sum_{t=1}^{k}\frac{\gamma_t^2\sigma^2}{2(1 - L\gamma_t)} \tag{5}$$

Divide bothsides by $\sum \gamma_t \Rightarrow$:

$$\frac{1}{\sum_{t=1}^{k}\gamma_t}\sum_{t=1}^{k} \gamma_t \mathbb{E}\left[f(x_{t+1}) - f(x)\right] \le \frac{1}{\sum_{t=1}^{k}\gamma_t}\left(\frac{1}{2}\|x - x_1\|^2 + \sum_{t=1}^{k}\frac{\gamma_t^2\sigma^2}{2(1 - L\gamma_t)}\right) \tag{6}$$

And, as before, define the output as weighted avg:

$$\overline{x}_{t+1} := \frac{\sum_{t=1}^{k}\gamma_t x_{t+1}}{\sum_{t=1}^{k}\gamma_t}$$

And (6) becomes:

$$\mathbb{E}\left[f(\overline{x}_{t+1}) - f(x)\right] \le \frac{1}{\sum_{t=1}^{k}\gamma_t}\left(\frac{1}{2}\|x - x_1\|^2 + \sum_{t=1}^{k}\frac{\gamma_t^2\sigma^2}{2(1 - L\gamma_t)}\right) \tag{7}$$

If $\gamma_t \le 1/2L$: $1 - L\gamma_t \ge 1 - 1/2 = 1/2$. And (7) becomes[2]

$$\mathbb{E}\left[f(\overline{x}_{t+1}) - f(x)\right] \le \frac{1}{\sum_{t=1}^{k}\gamma_t}\left(\frac{1}{2}D_X^2 + \sigma^2\sum_{t=1}^{k}\gamma_t^2\right) \tag{8}$$

Suppose $\gamma_t \equiv \gamma \le 1/2L$ for $t = 1, ..., k$. $\Rightarrow$

$$\mathbb{E}\left[f(\overline{x}_{t+1}) - f(x)\right] \le \frac{D_X^2}{2\gamma k} + \gamma\sigma^2 \tag{9}$$

Now we want to minimize RHS such that $\gamma \le 1/2L$ (an extra constraint). The solution is $\gamma^* = \min\left\{\frac{D_X}{\sigma\sqrt{2k}}, \frac{1}{2L}\right\}$. Insert γ^* into (9) in place of γ:

$$\begin{aligned}\mathbb{E}\left[f(\overline{x}_{t+1}) - f(x)\right] &\le \frac{D_X^2}{2k}\max\left\{\frac{\sigma\sqrt{2k}}{D_X}, 2L\right\} + \sigma^2\min\left\{\frac{D_X}{\sigma\sqrt{2k}}, \frac{1}{2L}\right\} \\ &\le \max\left\{\frac{D_X\sigma}{\sqrt{2k}}, \frac{LD_X^2}{k}\right\} + \frac{D_X\sigma}{\sqrt{2k}} \\ &\le \frac{LD_X^2}{k} + \frac{2D_X\sigma}{\sqrt{2k}}\end{aligned} \tag{10}$$

Let $x = x^*$, if we want to control the error such that $\mathbb{E}\left[f(\overline{x}_{t+1}) - f(x^*)\right] \le \epsilon$, we can solve k inversely:

$$k \ge \frac{2LD_X^2}{\epsilon} + \frac{8D_X^2\sigma^2}{\epsilon^2}$$

[2] $D_X :=$ diameter of X.

Remark.1 $f(x) = \mathbb{E}\left[F(x, \xi)|\xi\right]$, SGD is nearly an optimal algorithm.

Remark.2 $f(x) = \sum_{i=1}^{d} f_i(x)$ is a deterministic problem but *can be treated as an expectation problem.* We can improve the rate of convergence in terms of the dependence on ϵ. But the convergence depends on d.