10.5 Analysis of Stochastic Gradient Descent

SGD

$\Gamma : \mathcal{F} \to \mathbb{R}$

continuously differentiable

$(\xi_t : t \geq 1)$ independent

$f_t \to f^*$

$g(f_t, \xi_t)$ – "stochastic gradient"

SGD: $f_{t+1} = f_t - \alpha \underbrace{g(f_t, \xi_t)}_{\nabla\Gamma(f_t) \text{ for gradient alg.}}$ (f_0 given)

Example

training data

$$\Gamma(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f, z_i) \qquad \nabla\Gamma(f) = \frac{1}{n}\sum_{i=1}^{n} \nabla \ell(f, z_i)$$

$\xi_t \sim$ unif. over $[n] = \{1, \cdots, n\}$

$$g(f, \xi_t) = \nabla \ell(f, z_{\xi_t})$$

or

for some $k \geq 1$, ξ_t = a random subset of $[n]$ with $|\xi_t| = k$

$$g(f, \xi_t) = \frac{1}{k} \sum_{i \in \xi_t} \nabla \ell(f, z_i)$$

Assumption 10.1

(i) For some $\mu > 0$ so

$$\langle \nabla\Gamma(f_t), E_{\xi_t} g(f_t, \xi_t) \rangle \geq \mu \|\nabla\Gamma(f_t)\|^2$$

(ii) $$E_{\xi_t}\left[\|g(f_t, \xi_t)\|^2\right] \leq B + B_G \|\nabla\Gamma(f_t)\|^2$$

Thm 10.7 Assume Assump 10.1 and Γ is M-smooth m-strongly convex

$M \geq m > 0$

Then if SGD is run with $0 < \alpha \leq \frac{\mu}{M B_G}$ then

$$E\,\Gamma(f_t) - \Gamma^* \leq \underbrace{\frac{\alpha M B}{2 m \mu}}_{\text{steady state bound}} + (1 - \alpha m \mu)^{t-1}\left(\Gamma(f_1) - \Gamma^* - \frac{\alpha M B}{2 m \mu}\right)$$

$$\Gamma(f_{t+1}) - \Gamma(f_t) \leq \langle \nabla\Gamma(f_t), f_{t+1} - f_t \rangle + \frac{M}{2}\|f_{t+1} - f_t\|^2$$

smoothness

$$= -\alpha \langle \nabla\Gamma(f_t), g_t \rangle + \frac{\alpha^2 M}{2}\|g_t\|^2$$

$$E_\xi \Gamma(f_{t+1}) - \Gamma(f_t) \leq -\alpha\mu \|\nabla\Gamma(f_t)\|^2 + \frac{\alpha^2 M}{2}\left[B + B_G \|\nabla\Gamma(f_t)\|^2\right]$$

$$= -\alpha\left(\mu - \frac{\alpha M B_G}{2}\right)\|\nabla\Gamma(f_t)\|^2 + \frac{\alpha^2 M B}{2}$$

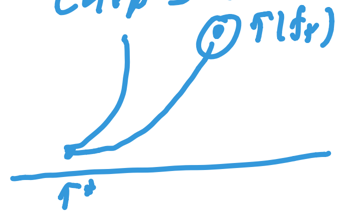
{by choice of α

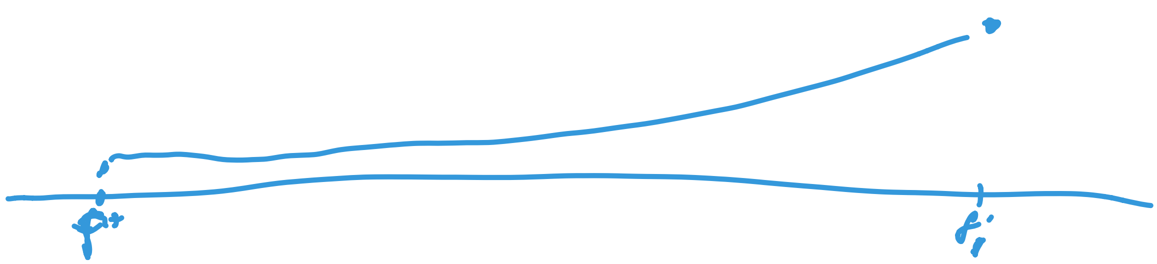
$$E_\xi \Gamma(f_{t+1}) - \Gamma(f_t) \leq -\frac{\alpha\mu}{2}\|\nabla\Gamma(f_t)\|^2 + \frac{\alpha^2 M B}{2}$$

By m-strong convexity

$$\Gamma(f_t) - \Gamma^* \leq \frac{\|\nabla\Gamma(f_t)\|^2}{2m}$$

Chap 3 Lemma 3.2(5)





so

$$E_\xi F(f_{t+1}) - f(\xi_t) \le -\alpha \mu m \left(F(f_t) - F^*\right) + \frac{\alpha^2 MB}{2}$$

Next take total expectation on both sides

$$E F(f_{t+1}) - E F(f_t) \le -\alpha \mu m \left(E F(f_t) - F^*\right) + \frac{\alpha^2 MB}{2}$$

$$E F(f_{t+1}) - F^* \le E F(f_t) - \alpha \mu m \left(E F(f_t) - F^*\right) + \frac{\alpha^2 MB}{2} - F^*$$

$$\underbrace{E F(f_{t+1}) - F^* - \frac{\alpha MB}{2m\mu}} = (1 - \alpha m\mu) \underbrace{\left(E F(f_t) - F^* - \frac{\alpha MB}{2m\mu}\right)}$$

so

$$E F(f_t) - F^* - \frac{\alpha MB}{2m\mu} \le (1 - \alpha m\mu)^{t-1} \left(E F(f_1) - F^* - \frac{\alpha MB}{2m\mu}\right)$$

Thm 10.8 Instead let $\alpha_t = \frac{c}{\gamma + t}$

$c > \frac{1}{m\mu}$, γ so large that $\alpha_1 \le \frac{\mu}{MB_G}$

then $$E F(f_{t+1}) - F \le \frac{\nu}{\gamma + t}$$

where $\nu \triangleq \max\{ \quad , \quad \}$

Proof Recall

$$E\Gamma(f_{t+1}) - E\Gamma(f_t) \leq -\alpha\mu m\left(E\Gamma(f_t) - \Gamma^*\right) + \frac{\alpha^2 MB}{2}$$

Add $E\Gamma(f_t) - \Gamma^*$ to both sides to get:

$$E\Gamma(f_{t+1}) - \Gamma^* \leq (1-\alpha m\mu)\left(E\Gamma(f_t) - \Gamma^*\right) + \frac{\alpha^2 MB}{2}$$

(which also appears p. 131 in notes)

with α replaced by α_t to yield

$$\begin{aligned}
\mathbf{E}\Gamma(f_{t+1}) - \Gamma^* &\leq (1 - \alpha_t\sigma\mu)\left(\mathbf{E}\Gamma(f_t) - \Gamma^*\right) + \frac{\alpha^2\beta M}{2}, \\
&= \left(1 - \frac{c\sigma\mu}{\gamma + t}\right)\left(\mathbf{E}\Gamma(f_t) - \Gamma^*\right) + \frac{c^2\beta M}{2(\gamma + t)^2} \\
&\leq \left(1 - \frac{c\sigma\mu}{\gamma + t}\right)\left(\mathbf{E}\Gamma(f_t) - \Gamma^*\right) + \frac{\nu(c\sigma\mu - 1)}{(\gamma + t)^2},
\end{aligned} \tag{10.52}$$

where the last step uses the fact $\frac{c^2\beta M}{2} \leq \nu(c\sigma\mu - 1)$. Finally, (10.50) holds for $t = 1$ by the choice of ν, and it follows for all $t \geq 1$ from (10.52) by induction on t. $\square$