Convergence of SGD

Notation

For $w \in \mathbb{R}^d$ we write $f(w) \in \mathbb{R}$. The Euclidean norm is $\|\cdot\|$ and the gradient of f is ∇f.

Lemma 1
If f is L-smooth, i.e. $\|\nabla f(w) - \nabla f(w')\| \le L\|w - w'\|$ for all $w, w' \in \mathbb{R}^d$, then

$$f(w') \;\le\; f(w) + \langle \nabla f(w), w' - w\rangle + \frac{L}{2}\|w' - w\|^2.$$

Proof. Define $g(t) = f\big(w + t(w' - w)\big)$ for $t \in [0, 1]$.

Remark. *$g(0) = f(w)$ and $g(1) = f(w')$, so $f(w') - f(w) = \int_0^1 g'(t)\,dt$.*

The chain rule gives $g'(t) = \langle \nabla f(w + t(w' - w)),\ w' - w\rangle$. Insert and subtract $\nabla f(w)$:

$$g'(t) = \langle \nabla f(w), w' - w\rangle + \langle \nabla f(w + t(w' - w)) - \nabla f(w), w' - w\rangle.$$

By Cauchy–Schwarz and L-smoothness, $|\langle \nabla f(\cdot) - \nabla f(w), w' - w\rangle| \le Lt\|w' - w\|^2$. Hence

$$f(w') - f(w) \le \langle \nabla f(w), w' - w\rangle + \int_0^1 Lt\|w' - w\|^2\,dt = \langle \nabla f(w), w' - w\rangle + \frac{L}{2}\|w' - w\|^2.$$

$\square$

Set-up for SGD

Let $\{w_t\}_{t\geq 0} \subset \mathbb{R}^d$ be the iterates of

$$w_{t+1} \;=\; w_t - \alpha_t g_t, \qquad g_t \text{ a stochastic gradient sample.}$$

Definition 2 (1) f is L-smooth.

(2) $\mathbb{E}[g_t \mid w_t] = \nabla f(w_t)$ (unbiasedness).

(3) $\mathbb{E}\big[\|g_t - \nabla f(w_t)\|^2\big] \leq \sigma^2$ (bounded variance).

(4) Stepsizes: $\sum_{t=1}^{\infty} \alpha_t = +\infty$ and $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$.

Theorem 3

Under Definition 2,

$$\liminf_{t\to\infty} \mathbb{E}\big[\|\nabla f(w_t)\|^2\big] = 0.$$

Proof. Lemma 1 with $w' = w_{t+1} = w_t - \alpha_t g_t$ gives

$$f(w_{t+1}) \leq f(w_t) - \alpha_t \langle \nabla f(w_t), g_t \rangle + \frac{L\alpha_t^2}{2} \|g_t\|^2.$$

Taking conditional expectation and using (2)–(3),

$$\mathbb{E}\big[f(w_{t+1}) \mid w_t\big] \leq f(w_t) - \alpha_t \|\nabla f(w_t)\|^2 + \frac{L\alpha_t^2}{2} \big(\|\nabla f(w_t)\|^2 + \sigma^2\big).$$

Hence

$$\mathbb{E}\big[f(w_{t+1})\big] \leq \mathbb{E}\big[f(w_t)\big] - \alpha_t \Big(1 - \frac{L\alpha_t}{2}\Big) \mathbb{E}\big[\|\nabla f(w_t)\|^2\big] + \frac{L\alpha_t^2 \sigma^2}{2}.$$

Choose t such that $1 - \frac{L\alpha_t}{2} \geq \frac{1}{2}$. Summing from $t = 1$ to T (telescoping) yields

$$\sum_{t=1}^{T} \alpha_t \, \mathbb{E}\big[\|\nabla f(w_t)\|^2\big] \leq 2\big(\mathbb{E}[f(w_1)] - f^\star\big) + L\sigma^2 \sum_{t=1}^{T} \alpha_t^2,$$

where $f^\star := \inf_w f(w)$. Divide by $\sum_{t=1}^{T} \alpha_t$ and send $T \to \infty$; by (4) the denominator diverges while the $\sum \alpha_t^2$ term stays finite, so the limit-inferior of the expected gradient norm must be 0. $\square$