Lecture A: A brief overview of convergence of gradient descent

August 4, 2021

Lecturer: Lei Wu *Scribe: Lei Wu*

Consider the problem of minimizing

$$\min_{\theta} \hat{\mathcal{R}}_n(\theta).$$

The gradient descent (GD) iterates as follows

$$\theta_{t+1} = \theta_t - \eta_t \nabla \hat{\mathcal{R}}_n(\theta_t),$$

where η_t is the learning rate. When $\eta_t \to 0$, the GD becomes the GD flow:

$$\frac{\mathrm{d}\theta_t}{\mathrm{d}t} = -\nabla \hat{\mathcal{R}}_n(\theta_t).$$

Theorem 0.1 (Non-convex)**.** *For any $t > 0$,*

$$\min_{s\in[0,t]} \|\nabla \hat{\mathcal{R}}_n(\theta_s)\| \leq \sqrt{\frac{\hat{\mathcal{R}}_n(\theta_0) - \inf_\theta \hat{\mathcal{R}}_n(\theta)}{t}}.$$

Proof. The energy dissipation satisfies

$$\frac{\mathrm{d}\hat{\mathcal{R}}_n(\theta_t)}{\mathrm{d}t} = -\|\nabla \hat{\mathcal{R}}_n(\theta_t)\|_2^2.$$

Hence,

$$\hat{\mathcal{R}}_n(\theta_0) - \hat{\mathcal{R}}_n(\theta_t) = \int_0^t \|\nabla \hat{\mathcal{R}}_n(\theta_s)\|_2^2 \,\mathrm{d}s \geq t \min_{s\in[0,t]} \|\nabla \hat{\mathcal{R}}_n(\theta_s)\|^2.$$

□

The above theorem shows that GD will converge to a stationary point, which is the best we can expect for general non-convex problem. Next, we prove that GD will converge to a global minima, if the objective function is convex.

Theorem 0.2. *Assume that $\hat{\mathcal{R}}_n$ is convex and the minimizer is given by θ^* with $\|\theta^*\|_2 < \infty$. Then, we have*

$$\hat{\mathcal{R}}_n(\theta_t) - \hat{\mathcal{R}}_n(\theta^*) \leq \frac{\|\theta^* - \theta_0\|_2^2}{2t}.$$

Proof. For any $\bar{\theta}$, define

$$J(t) = t(\hat{\mathcal{R}}_n(\theta_t) - \hat{\mathcal{R}}_n(\bar{\theta})) + \frac{1}{2}\|\theta_t - \bar{\theta}\|_2^2.$$

Using the convexity, we have

$$\frac{\mathrm{d}J(t)}{\mathrm{d}t} = \hat{\mathcal{R}}_n(\theta_t) - \hat{\mathcal{R}}_n(\bar{\theta}) - t\|\hat{\mathcal{R}}_n(\theta_r)\|_2^2 + \langle \theta_t - \bar{\theta}, -\nabla \hat{\mathcal{R}}_n(\theta_t)\rangle$$

$$
\begin{aligned}
&= -t\|\hat{\mathcal{R}}_n(\theta_r)\|_2^2 - \Big(\hat{\mathcal{R}}_n(\bar{\theta}) - \hat{\mathcal{R}}_n(\theta_t) - \langle \bar{\theta} - \theta_t, \nabla \hat{\mathcal{R}}_n(\theta_t)\rangle\Big) \\
&\leq 0.
\end{aligned}
$$

Hence, $J(t) \leq J(0)$, i.e.,

$$
t(\hat{\mathcal{R}}_n(\theta_t) - \hat{\mathcal{R}}_n(\bar{\theta})) + \frac{1}{2}\|\theta_t - \bar{\theta}\|_2^2 \leq \frac{1}{2}\|\theta_0 - \bar{\theta}\|_2^2
$$

Taking $\bar{\theta} = \theta^*$ completes the proof. □

A natural question is that: Can we prove the converge to global minima for non-convex problem? This problem often strongly depends on the specific model. There exists a general condition as follows.

Definition 0.3 (Polyak-Lojasiewicz (PL) condition)**.** $\hat{\mathcal{R}}_n$ is said to satisfy the PL condition if

$$
\|\nabla \hat{\mathcal{R}}_n(\theta)\|_2 \geq C(\hat{\mathcal{R}}_n(\theta) - \inf_{\theta} \hat{\mathcal{R}}_n(\theta)).
$$

Theorem 0.4. *Under the PL condition, we have*

$$
\hat{\mathcal{R}}_n(\theta_t) - \inf_{\theta} \hat{\mathcal{R}}_n(\theta) \leq e^{-Ct}(\hat{\mathcal{R}}_n(\theta_0) - \inf_{\theta} \hat{\mathcal{R}}_n(\theta)).
$$

Proof.

$$
\frac{\mathrm{d}\hat{\mathcal{R}}_n(\theta_t)}{\mathrm{d}t} = -\|\nabla \hat{\mathcal{R}}_n(\theta)\|_2^2 \leq -C(\hat{\mathcal{R}}_n(\theta) - \inf_{\theta} \hat{\mathcal{R}}_n(\theta)) \tag{0.1}
$$

Let $\Delta_t = \hat{\mathcal{R}}_n(\theta_t) - \inf_\theta \hat{\mathcal{R}}_n(\theta)$. Then, $\dot{\Delta}_t \leq -C\Delta_t$. Hence, $\Delta_t \leq e^{-Ct}\Delta_0$. □

Remark 0.5. Strongly convex functions satisfy the PL condition.