So far: Convergence properties on cvx problems

This lecture: Some guarantees for some nonconvex problems

Q: What can we say about general non-cvx problems?

A: Not much because $f(w)$ can have: • exp. num of local mins
• " " " saddle points

But, we can show convergence to critical point, or add structure and prove bounds for some non-cvx losses

Q: Why can't we show global convergence?

A: In general $\min_w f(w)$ can encode NP-Hard problems.

What about <u>local convergence</u>?

SGD on <u>smooth non-cvx</u> fns:

<u>Reminder:</u>

f is β-smooth if

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x-y\|, \quad \forall x, y \in \mathbb{R}^d$$

This implies:

$$|f(x) - f(y) - \langle \nabla f(y), x-y \rangle| \leq \frac{\beta}{2} \|x-y\|^2 \quad \forall x, y \in \mathbb{R}^d$$

Let $x = w_{k+1}$, $y = w_k$

$$w_{k+1} = w_k - \gamma \nabla f_{S_k}(w_k)$$

Then smoothness implies

$$f(w_{k+1}) - f(w_k) - \langle \nabla f(w_k), w_{k+1} - w_k \rangle \leq \frac{\beta}{2} \| \gamma \nabla f_{S_k}(w_k) \|^2$$

$$\Rightarrow f_{k+1} \leq f_k - \gamma \langle \nabla f(w_k), \nabla f_{S_k}(w_k) \rangle + \frac{\beta \gamma^2}{2} \| \nabla f_{S_k}(w_k) \|^2$$

$$\Rightarrow \mathbb{E} f_{k+1} \leq \mathbb{E} f_k - \gamma \, \mathbb{E} \| \nabla f(w_k) \|^2 + \frac{\beta \gamma^2}{2} \underbrace{\| \nabla f_{S_k}(w_k) \|^2}_{\leq M^2}$$

$$\leq \mathbb{E} f_k - \gamma \, \mathbb{E} \| \nabla f(w_k) \|^2 + \frac{\beta \gamma^2 M^2}{2}$$

$$\Rightarrow \gamma \, \mathbb{E} \| \nabla f(w_k) \|^2 \leq \mathbb{E}(f_k - f_{k+1}) + \frac{\beta \gamma^2 M^2}{2}$$

$$\Rightarrow \mathbb{E} \| \nabla f(w_k) \|^2 \leq \frac{\mathbb{E}(f_k - f_{k+1})}{\gamma} + \frac{\beta \gamma M^2}{2}$$

Lets now sum for T iterations

$$\mathbb{E}\|\nabla f(w_1)\|^2 \le \frac{\mathbb{E} f_1 - f_2}{\gamma} + \beta\gamma M^2$$

$$\mathbb{E}\|\nabla f(w_2)\|^2 \le \frac{\mathbb{E} f_2 - f_3}{\gamma} + \beta\gamma M^2$$

$$\vdots$$

$$\sum_{k=1}^{T} \mathbb{E}\|\nabla f(w_k)\|^2 \le \frac{\mathbb{E} f_1 - f^*}{\gamma} + T\cdot\beta\cdot\gamma M^2$$

$$\Rightarrow T \cdot \min_{t=1:T} \mathbb{E}\|\nabla f(w_t)\|^2 \le \frac{\mathbb{E} f_1 - f^*}{\gamma} + T\beta\gamma M^2$$

$$\Rightarrow \min_{t=1:T} \mathbb{E}\|\nabla f(w_t)\|^2 \le \frac{\mathbb{E} f_1 - f^*}{T\gamma} + \beta\gamma M^2$$

Lets assume $f_1 - f^* \le D$

then set $\gamma = \sqrt{\frac{D}{\beta M^2 T}}$ to get:

$$\min_t \mathbb{E}\|\nabla f(w_t)\|^2 \le 2\sqrt{\frac{D\beta M^2}{T}}$$

The above implies that no matter what $f(w)$ looks like, if it is smooth it can reach an "approximate" critical point in $O(1/\sqrt{T})$ iterations.

This rate seems too slow to explain practial perf. of SGD.

What if we add a bit more structure

Polyak-Łojasiewicz (PL) - condition:

A function is μ-PL if

$$\|\nabla f(x)\|^2 \geq \mu(f(x) - f^*)$$

in combination with β- smooth

$$\Rightarrow \quad \beta^2 \|x - \Pi_{x^*}(x)\|^2 \geq \mu(f(x) - f^*)$$

That is, the furter we move from the opt the the more the suboptimality increases.

Remark: Not all PL fns are cvx.
If a cvx fn is PL, then it's strongly-cvx.

Remark: All local minima are global.
(This happens in some NN cases)

SGD on PL fns:

From previous analysis:

$$\mathbb{E} f_{k+1} \le \mathbb{E} f_k - \gamma \, \mathbb{E} \|\nabla f(\omega_k)\|^2 + \beta \gamma^2 M^2$$

$$\le \mathbb{E} f_k - \mu\gamma (\mathbb{E} f_k - f^*) + \beta \gamma^2 M^2$$

Subtract f^* from both sides

$$\Rightarrow \quad \mathbb{E} f_{k+1} - f^* \le (1 - \mu\gamma) \mathbb{E}(f_k - f^*) + \beta \gamma^2 M^2$$

$$\le (1-\mu\gamma)^{k+1} (f_0 - f^*) + \sum_{i=0}^{k} (1-\mu\gamma)^i \cdot \beta \gamma^2 M^2$$

$$\leq (1-\mu\gamma)^{K+1} D + \frac{B\gamma^2 M^2}{\mu\gamma}$$

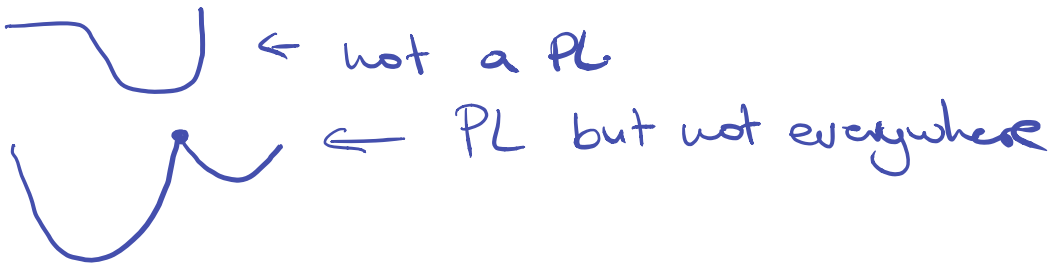
$$= \underbrace{(1-\mu\gamma)^{K+1} D}_{\varepsilon/2} + \underbrace{\frac{B\gamma M^2}{\mu}}_{\varepsilon/2}$$

$$\gamma = \frac{\varepsilon \cdot \mu}{B M^2} \qquad T = O\left(\frac{M^2 \cdot B}{\mu^2} \frac{1}{\varepsilon} \log \frac{D}{\varepsilon}\right)$$

Much faster conv. rate.

Remark: PL is a useful property but does not account for saddle points

e.g.

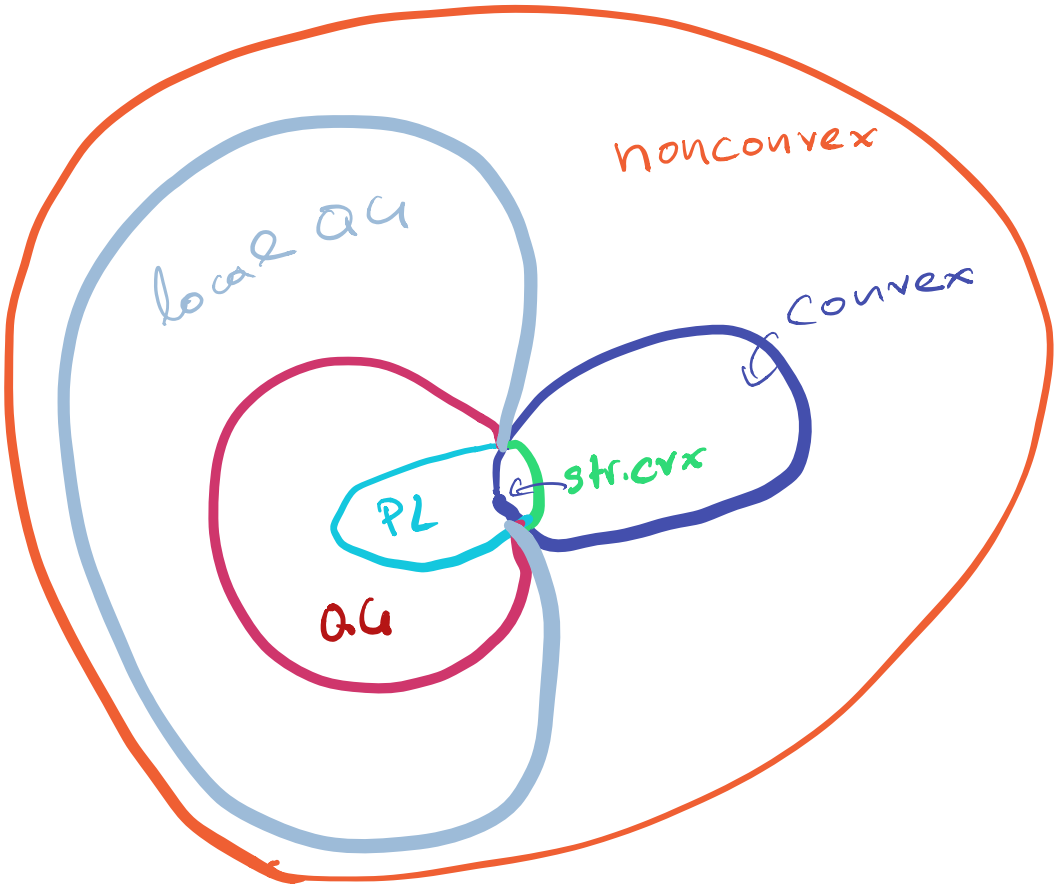


← not a PL

← PL but not everywhere

A more general fn class : Quadratic Growth

$$f(w) - f^* \geq \mu/2 \, \| w - \Pi_{W^*}(w) \|^2$$



Unfortunately no conv. guarantees known for QG.

Also unclear how relevant it is in practice

Several open Qs:

- What are interesting non-cvx fns where we can show convergence?
- Do nns have as many local min as global min?
- What is the suboptimality of local min on real problems.

Next time:

- How to really choose stepsize?
- What happens for non-diff. functions?
- What to do when we have constraints?