1) Q Function Approximation

$$\{(\underbrace{s_i, a_i}_{\text{rollout}}, \underbrace{y_i}_{\substack{\text{MC} \\ \text{TD} \\ \text{Bellman opt.}}})\}_{i=1}^{N}$$

$$\hat{Q} = \arg\min_{\theta} \sum_{i=1}^{N} \underbrace{(Q_\theta(s_i, a_i) - y_i)^2}$$

$$\mathcal{Q} = \{ Q_\theta : \theta \in \mathbb{R}^d \}$$

Incremental Updates: gradient-based method

$$\boxed{\nabla_\theta \left[(Q_\theta(s_i, a_i), y_i)^2 \right]} = 2 \underbrace{(Q_\theta(s_i, a_i) - y_i)}_{\text{error}} \underbrace{\nabla_\theta Q_\theta(s_i, a_i)}$$

update

$$\theta \leftarrow \theta - \alpha \left(Q_\theta(s_i, a_i) - \underbrace{y_i}\right) \nabla_\theta Q_\theta(s_i, a_i)$$

$$y_i = r_i + \gamma \underbrace{Q_\theta(s_{i+1}, a_{i+1})}$$

How to choose (s_i, a_i, y_i) to update with respect to:

1) online GD uses (s_t, a_t, y_t)
2) "experience replay": store incoming data
   resample $\{(s_i, a_i, y_i)\}_{i=1}^{N}$ randomly from stored data
   for $i = 1, \dots, N$

$$\theta \leftarrow \theta - \alpha \nabla_\theta \ell (Q_\theta(s_i, a_i) - y_i)^2$$

2) Optimization & Gradient Ascent

Goal: find π^* (approximately). Why bother with Q^*?

suppose parametrized policy π_θ

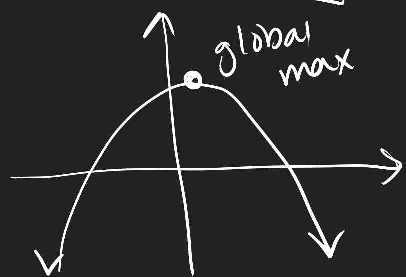
objective function: $\underbrace{J(\theta)} = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, P, \pi_\theta, s_0\right]$

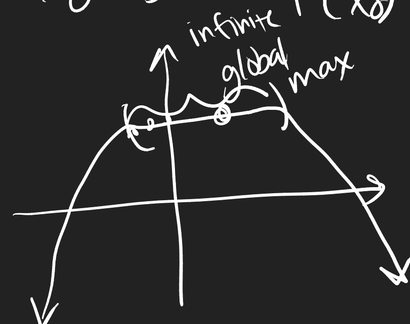
Maxima & Minima: $f(x): \mathbb{R}^d \to \mathbb{R}$

A global max is x_0 s.t. $f(x_0) \geq f(x)$ $\forall x \in \mathbb{R}^d$

A local max is x_0 s.t. $f(x_0) \geq f(x)$ $\forall \|x - x_0\| \leq \varepsilon$

$\exists\, \varepsilon > 0$

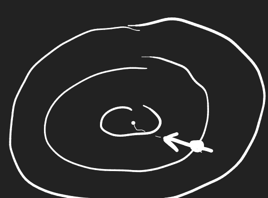




An ascent direction is $v \in \mathbb{R}^d$ s.t.

$f(x + \alpha v) > f(x)$ for some $\alpha > 0$

If f is differentiable, the $\nabla f(x)$ is the steepest ascent direction



Gradient Ascent:

initial x_0

for $t = 0, 1, \ldots$

$\underline{x_{t+1} = x_t + \alpha \nabla f(x_t)}$

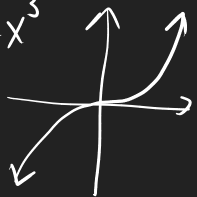
α ← step size

$\underbrace{f(x)}_{\max f(x)} \approx \underbrace{f(x_t) + \nabla f(x_t)^T (x - x_t)}$, when x & x_t close

$x - x_t \propto \nabla f(x_t)$ $\qquad$ α keeping close

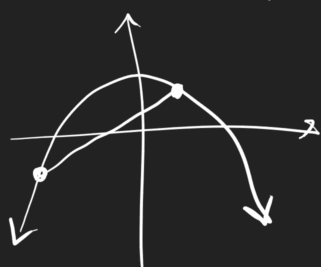
<u>Fact</u> $\nabla f(x_0) = 0 \Leftarrow x_0$ local max

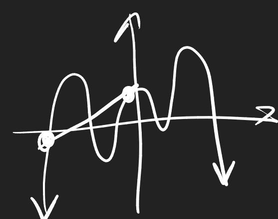
$f(x) = x^3$



<u>Critical Points:</u> $\forall x$ s.t. $\nabla f(x) = 0$
- local / global max/min
- saddle points

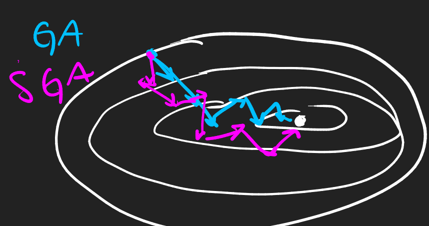
<u>Concave</u> function has critical point(s) which are global maxima





3) Stochastic Gradien Ascent

Instead $\nabla f(x_t)$, we have g_t s.t. $\mathbb{E}[g_t] = \nabla f(x_t)$



<u>Alg: SGA</u>
init x_0
for $t = 0, 1, \ldots$
$\quad x_{t+1} = x_t + \alpha g_t$

<u>ex</u>: SGD for ERM

$$\min_\theta \underbrace{\frac{1}{N} \sum_{i=1}^{N} \ell(f_\theta(x_i), y_i)}$$

$$\underline{g_t} = \nabla_{\theta_t} \ell(f_{\theta_t}(x_i, y_i))$$

$x_i, y_i \sim$ uniforly Sampled

$$\mathbb{E}[g_t] = \sum_{i=1}^{n} \nabla_{\theta_t} \ell(f_{\theta_t}(x_i), y_i) \cdot \frac{1}{N}$$

Theorem: Assume:

- 1) $f(x)$ is β smooth $\|\nabla f(x) - \nabla f(x')\|_2 \leq \beta \|x - x'\|_2$
- 2) $\sup_x |f(x)| \leq M$
- 3) $\mathbb{E}[g(x)] = \nabla f(x)$
- 4) $\mathbb{E}[\|g(x)\|_2^2] \leq \sigma^2$

Then SGA with $g_t = g(x_t)$ converges

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(x_t)\|_2\right] \lesssim \sqrt{\frac{\beta\sigma^2 M}{T}} \qquad \alpha = \sqrt{\frac{M}{\beta\sigma^2 T}}$$

Question can a sampled trajectory $\tau = (s_0, a_0, \dots)$ directly give us estimate of $\nabla J(\theta)$?

A: No! not knowing transition P or reward r is like not knowing loss function in SL

$$J(\theta) = \mathbb{E}_{w \sim N(0,1)}\left[s_1^2 \mid s_1 = f(s_0, a_0, w),\ a_0 = \pi_\theta(s_0)\right]$$

$$\nabla J(\theta) = \nabla_\theta \mathbb{E}_w\left[f(s_0, \pi_\theta(s_0), w)^2\right]$$

$$= \mathbb{E}_w\left[\nabla_\theta f(s_0, \pi_\theta(s_0), w)^2\right] \neq \mathbb{E}[s_1^2]$$

one trajectory: (s_0, a_0, s_1)

Next Lecture: Derivative-Free Optimization

can't access $\nabla f(x)$ or even estimate

but can access $f(x)$