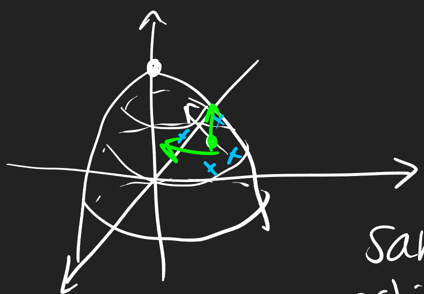
1) Derivative-Free Optimization.

find maximum of $f(\theta): \mathbb{R}^d \to \mathbb{R}$

only using function evaluation — $f(\theta)$, not $\nabla f(\theta)$

<u>Goal</u>: find a ascent direction



sample nearby points

see which lead to increase in f

We focus methods that use samples to construct a gradient estimate

A) Random Search

Recall finite difference approximation

1D) $f'(\theta) \approx \dfrac{f(x+\delta)-f(x-\delta)}{2\delta}$

vector fn) $\nabla f(\theta)^T v \approx \dfrac{f(x+\delta v) - f(x-\delta v)}{2\delta}$

"directional derivative"

<u>Alg: Random Search</u>

initial θ_0

for $t=0,1,\ldots$

sample $v_1, \ldots, v_N \overset{iid}{\sim} \mathcal{N}(0, I)$

update $\theta_{t+1} = \theta_t + \alpha \cdot \frac{1}{2\delta N} \sum_{k=1}^{N} \big(f(\theta_t+\delta v_k) - f(\theta_t - \delta v_k)\big) v_k$

We can undstand this in terms of SGA, $\theta_t + \alpha g_t$, $\mathbb{E}[g_t] = \nabla f(\theta)$

$$\mathbb{E}_{v_k}\left[\frac{1}{2\delta N}\sum_{k=1}^{N}\left[f(\theta_t+\delta v_k) - f(\theta_t - \delta v_k)\right] v_k\right]$$

$$\approx \mathbb{E}_{v_k}\left[\frac{1}{\cancel{2\delta} N}\sum_{k=1}^{N} \cancel{2\delta} \cdot \left(\nabla f(\theta_t)^T v_k\right) \cdot v_k\right]$$

$$= \frac{1}{N}\sum_{k=1}^{N} \mathbb{E}_{v_k}\left[v_k v_k^T\right]\nabla f(\theta_t)$$

$$= \frac{1}{N}\sum_{k=1}^{N} \nabla f(\theta_t) = \nabla f(\theta_t)$$

This method samples in parameter space: θ

B) Importance weighting

Distribution trick:

$$f(\theta) = \mathbb{E}_{x\sim P_\theta}[h(x)]$$

$\left[\begin{array}{l}\text{let } P_\theta = \mathbb{I}\{x=\theta\} \\ \text{let } h = f\end{array}\right.$

suppose another distribution $\rho(x)$ ← assume $<\infty$

$$\mathbb{E}_{x\sim P_\theta}[h(x)] = \sum_x h(x)\cdot P_\theta(x)\cdot\frac{\rho(x)}{\rho(x)} = \mathbb{E}_{x\sim\rho}\left[\frac{P_\theta(x)}{\rho(x)}h(x)\right]$$

write gradient:

$$\nabla_\theta f(\theta) = \nabla_\theta \mathbb{E}_{x\sim P_\theta}[h(x)] = \mathbb{E}_{x\sim\rho}\left[\frac{\nabla_\theta P_\theta(x)}{\rho(x)}h(x)\right]$$

pick $\rho(x) = P_\theta(x)$

$$\mathbb{E}_{x\sim P_\theta}\left[\frac{\nabla_\theta P_\theta(x)}{P_\theta(x)}h(x)\right] = \mathbb{E}_{x\sim P_\theta}\left[\nabla_\theta \log P_\theta(x)\, h(x)\right]$$

Alg:

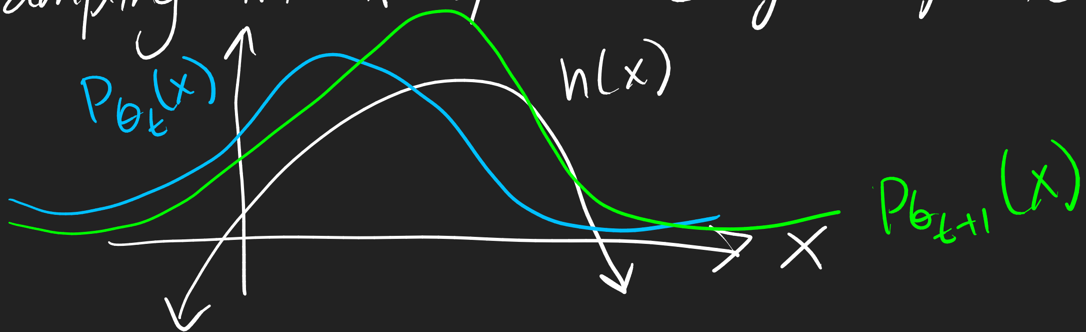
initialize θ_0

for $t = 0, 1, \ldots$

sample $x_k \sim P_{\theta_t}$ for $k = 1, \ldots, N$

$$\theta_{t+1} = \theta_t + \frac{\alpha}{N}\sum_{k=1}^{N}\nabla_{\theta_t}\log P_{\theta_t}(x_k)\,h(x_k)$$

Sampling in x-space (e.g. trajectory space)



2) Policy Optimization

Recall RL setting

- MDP $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, r, \gamma\}$ with P, r unknown
- parametrized policy π_θ $\quad \theta \in \mathbb{R}^d$
- objective function: $J(\theta) = \mathbb{E}_{s_0 \sim \mu_0}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,/\, P, r, \pi_\theta\right]$
- observe "rollout" of π_θ

  trajectory $\tau = (s_0, a_0, s_1, a_1, \dots)$

  rewards $(r_0, r_1, \dots)$

means we can "sample" $J(\theta) = \mathbb{E}_\tau[R(\tau)]$ $\quad R(\tau) = \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$

observe τ, $R(\tau)$

Meta Algorithm: Derivative Free SGA

initialize θ_0

for $t = 0, 1, \dots$

1) collect rollouts with θ_t
2) compute gradient estimate g_t of $\nabla J(\theta_t)$
3) $\theta_{t+1} = \theta_t + \alpha g_t$

The rest of lecture: 3 ways to estimate $\nabla J(\theta)$

Alg: Simple Random search

Based on "random search": 1) collect 2 rollouts

$\tau^+ : \pi_{\theta_t + \delta v}$ $v \sim N(0, I)$

$\tau^- : \pi_{\theta_t - \delta v}$

2) compute estimate: $g_t = \frac{1}{2\delta}\left(\underline{R}(\tau^+) - \underline{R}(\tau^-)\right)v$

3) Policy Gradient (PG) from Trajectories

(REINFORCE)

$\tau = (s_0, a_0, s_1, \ldots)$ $\quad \underline{\rho_\theta(\tau)} = \mu_\theta(s_0)\pi_\theta(a_0|s_0)P(s_1|s_0,a_0)\pi_\theta(a_1|s_1)\cdots$

$$J(\theta) = \mathop{\mathbb{E}}_{\tau \sim \rho_\theta}[R(\tau)]$$

x $\quad$ p_θ $\quad$ h(x)

Claim: 1) for $\tau \sim \rho_\theta$ (i.e. τ observed from "rolling out" π_θ)

2) $\rightarrow g = \sum_{t=0}^{\infty} \nabla_\theta[\log(\pi_\theta(a_t|s_t))]\underline{R(\tau)}$ is an unbiased estimate of $\nabla J(\theta)$

Why?

$$\nabla J(\theta) = \mathop{\mathbb{E}}_{\tau \sim \underline{\rho_\theta}}[\nabla_\theta[\log(\rho_\theta(\tau))]R(\tau)]$$

$$\nabla_\theta[\cancel{\log(\mu_\theta(s_0))} + \log(\pi_\theta(a_0|s_0)) + \cancel{\log(P(s_1|s_0,a_0)}) + \ldots]$$

don't depend on

4) PG with Value Functions

Claim: 1) $s, a \sim d_{\mu_0}^{\pi_\theta}$, 2) $g = \frac{1}{1-\gamma}\nabla_\theta[\log(\pi_\theta(a|s))][Q^{\pi_\theta}(s,a) - b(s)]$ is an unbiased estimate of $\nabla J(\theta)$

"baseline"