general-purpose algorithms

refs: Recht 2019 "A tour of ..."

Nesterov & Spokoiny 2017 "Random gradient-free ..."

Williams 1992 "Simple statistical ..."

- consider the unconstrained nonlinear program

$$\text{(NLP)} \quad \min \; c(u) \quad \text{s.t.} \; u \in \mathbb{R}^m$$

idea: iteratively choose u^+ such that $c(u^+) < c(u)$

* how to choose u^+ ?

- if c is continuously diff'able & we know $Dc(u)$, → $c \in C^1(\mathbb{R}^m, \mathbb{R})$

  → choose $u^+ = u - \gamma \cdot Dc(u)$

$$e_i = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow \text{i-th entry}$$

- if $c \in C^1$ but we don't know $\partial c(u)$,

  → approximate $[Dc(u)]_i \simeq g_i = \underbrace{\frac{1}{\Delta}\left(c(u + \Delta \cdot e_i) - c(u)\right)}_{\text{finite difference approx.}}$

  choose $u^+ = u - \gamma \cdot g$

- if $c \notin C^1$ or we can't approximate $\partial c(u)$

  → choose u^+ at random !?

- although random search may sound like a bad idea, it has solid theoretical foundations & empirical successes:
  - simulated annealing
  - genetic / swarm algorithms
- let's consider a specific class of randomized algorithms that leverage gradient-like information

→ instead of $\min c(u)$ s.t. $u \in \mathbb{R}^m$ consider $\min E[c(u)]$ s.t. $u \sim p$,

p is a probability distribution

* how does minimum / minimizer of — related to — ?

→ if "δ" distributions are allowed, then mini-ma/-mizers of 1st are minima/mizers of 2nd (and since $\int p = 1$ the 2nd can't have lower cost)

... but there are (uncountably) infinite "δ"-distributions ...

→ instead of $\min E[c(u)]$ s.t. $u \sim p$ consider $\min E[c(u)]$ s.t. $u \sim p_\theta$

where p_θ is a distribution parameterized by $\theta \in \Theta$

idea: "log-likelihood trick"

$$D_\theta E[c(u)] = D_\theta \int c(u)\, p_\theta(u)\, du$$ – def. of expectation

$$= \int c(u)\, D_\theta p_\theta(u)\, du$$ – assuming D & $\int$ commute

$$= \int c(u)\, D_\theta p_\theta(u) \frac{p_\theta(u)}{p_\theta(u)}\, du$$ – assuming $p_\theta(u) \neq 0$

$$= \int \left(c(u)\, D_\theta \log p_\theta(u)\right) p_\theta(u)\, du$$ – $D_x \log f(x) = \frac{D_x f(x)}{f(x)}$

$$= E\left[c(u)\, D_\theta \log p_\theta(u)\right]$$ – def. of expectation

* if we sample $u \sim p_\theta(u)$, evaluate $c(u)$, and average, we can compute the derivative $D_\theta E[c(u)]$ <u>without</u> derivative $D_u c$!

algo (REINFORCE) [Williams 1992]

1°. sample $\{u_n\}_{n=1}^N \sim p_\theta$

2°. update $\theta^+ = \theta - \gamma \cdot \underbrace{\frac{1}{N} \sum_n c(u_n) \cdot D_\theta \log p_\theta(u_n)}$

↳ $E[c(u)\, D_\theta \log p_\theta(u)] = \lim_{N \to \infty}$ by Central Limit Thm

∘ key facts: $E[c] \sim c$ when variance of p_θ is small

& $D_\theta E[c] \sim D_u c$ [Nesterov & Spokoiny 2017]

↳ REINFORCE is (stochastic) gradient descent on a smoothed cost function

[Recht 2019]

ex: "simple / pure random search"

note: specific parameterization $\theta \in \mathbb{R}^m$

- consider $p_\mu(u) = \rho(u - \mu)$ where ρ is
  - normal / Gaussian or - uniform on a sphere
- then REINFORCE simplifies to

$$\mu^+ = \mu - \gamma \cdot \underbrace{\frac{1}{\Delta}\left(c(\mu + \Delta u) - c(\mu)\right)}_{\text{unbiased estimate of directional derivative } D_\mu E[c(\mu)] \cdot u} \cdot u \quad \text{where } u \sim \rho$$