general-purpose algorithms

refs: Recht 2019 "A tour of ..."
Nesterov & Spokoiny 2017 "Random gradient-free ..."
Williams 1992 "Simple statistical ..."

- consider the unconstrained nonlinear program

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

want u^* s.t. $c(u^*) < c(u)$ for all $u \neq u^*$ $\underbrace{(\|u - u^*\| < \varepsilon)}_{\text{local minimum}}$

idea: iteratively choose u^+ such that $c(u^+) < c(u)$
and hope that $u \mapsto u^+ \mapsto u^{++} \mapsto \ldots \mapsto u^*$

* how to choose u^+?

→ two main strategies:

1°: randomize
$u^+ \sim p(u)$

2°: descend
$u^+ = u - \gamma \cdot Dc(u)$

eg
- simulated annealing
- genetic / swarm algo

eg
- gradient descent
- Newton-Raphson

+ easy implementation	+ easy[-ish] implementation
+ approximate global u^*	− approximate local u^*
− sample-inefficient	+ sample-efficient
+ need continuous c	− need differentiable c

- 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 — relate 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 $\min E[c(u)]$ consider $\min E[c(u)]$

s.t. $u \sim p$ s.t. $u \sim p_\theta$

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

* richness of parameterization trades off tractability with suboptimality

idea: "log-likelihood trick"

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

$$= \int c(u)\, D_\theta p_\theta(u)\, du \quad \text{– assuming } D \,\&\, \textstyle\int \text{ commute}$$

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

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

$$= E[c(u)\, D_\theta \log p_\theta(u)] \quad \text{– def. of expectation}$$

* why is this a useful result?

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

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

[Nesterov & Spokoiny 2017]

- key facts: when variance of p_θ is small: $E[c] \sim c$
  & $D_\theta E[c] \sim D_u c$

$\hookrightarrow$ REINFORCE is (stochastic) gradient descent on a smoothed cost function

[Recht 2019]

ex: "simple / pure random search"

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

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

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

* many "state-of-the-art" algorithms (eg policy gradient) are elaborations on the preceding

  → and the elaborations may not actually outperform the simple variants

  [Mania et al 2018 "Simple random search..."]