3/3: Zero-order methods

Announcements:
- HW 2 due Thurs (3/5)
- mid term survey next week
- HW 3 out Monday (3/9)

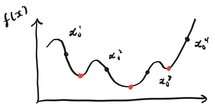
Last Time:
- convexity
- Sequential OPs

Today:
- optimizing w/o gradients
- cross-entropy method
- evolutionary strategies
- differential evolution

Motivation:

$\min_{v_0} \|p(T) - p_0\|$
s.t. $\dot{p}(t+1) = f(p_t)$

1. not going to assume access to gradient
2. multiple local optima



$\min_x f(x)$

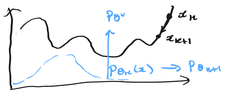
1. use multiple runs
2. inject noise into the optimization

Alg 1 (Stochastic Gradient Descent)
input: x_0, learning rate $\alpha > 0$
for $k = 0, 1, \dots, k_{max}$
$\varepsilon_k \sim \mathcal{N}(0, \sigma^2 I)$
$x_{k+1} \leftarrow x_k - \alpha(\nabla f(x_k) + \varepsilon_k)$

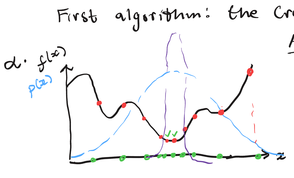
Key idea: proposal distribution

$\min_x f(x) \Rightarrow \{x_0, x_1, x_2, \dots, x^*\}$

$p_\theta(x)$ $\quad \min_\theta \mathbb{E}_{x \sim p_\theta}[f(x)]$



First algorithm: the Cross-Entropy Method (CEM)



ρ: 20% quantile
x_{elite}: best 2 of 10

Alg 1 (Cross-Entropy)
input: μ_0, Σ_0
for $k = 0, 1, \dots, k_{max}$
$x_k^i \sim \mathcal{N}(\mu_k, \Sigma_k)$
$f^i = f(x_k^i)$ ☆
$x_{elite} = \{x_k^i \mid f^i \geq q(f^i)\}$ (q: quantile)
$\mu_{k+1}, \Sigma_{k+1} = fit(x_{elite})$

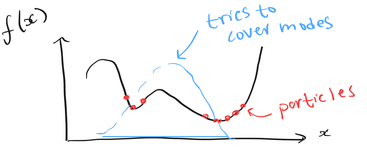
$\mu_{k+1} = \frac{1}{N_e} \sum_{x \in x_e} x$

$\Sigma_{k+1} = \frac{1}{N_e} \sum_{x \in x_e} (x - \mu_{k+1})(x - \mu_{k+1})^T$

Pros:
+ only require access to f
+ no matrix solves
+ robust to scaling of f
+ "embarrassingly parallel"

Cons:
- "throw away" gradient info
- unimodal

Neat extension: CMA-ES



Population methods: $X = \{x^1, \dots, x^N\}$
where each point minimizes $f(x)$

Stochastic → UKF (fit & refit dist p_θ)
population → particle filter (non-parametric)

One analogy: genetics/evolution

Alg 1 (Population method)
$X_0 \leftarrow$ initialize()
$f_0^i \leftarrow f(x_0^i)$ in X_0
for $k = 0, 1, \dots, k_{max}$
$\tilde{x}_k^i = mutate(X_k)$
$\tilde{f}_k^i = f(\tilde{x}_k^i)$ for $\tilde{x}_k^i \in \tilde{X}_k$
$X_{k+1} = select(X_k, f_k, \tilde{X}_k, \tilde{f}_k)$

Summary:
1. For non-convex functions, injecting noise can help find better local optima
2. Stochastic methods like CEM pose opt as $\min_\theta \mathbb{E}_{x \sim p_\theta}[f(x)]$
3. Other population methods come w/ few guarantees but explore effectively