Zeroth-order optimisation via GradientLess Descent

Jose Avilez

University of Waterloo

November 2020

The paper

This paper presentation is based on:
"Gradientless Descent: High-Dimensional Zeroth-Order Optimization"
By Daniel Golovin, John Karro, Greg Kochanski, Chansoo Lee, Xingyou Song, and Qiuyi Zhang from Google.

Zeroth-order optimisation via GradientLess Descent

Our set-up

Let $K \subseteq \mathbb{R}^n$ be compact. For a function $f : K \to \mathbb{R}$, we are interested in the basic optimisation problem:

$$x^* = \arg\min_{x \in K} f(x)$$

Under the following setting:

1. Convex or non-convex? Convex
2. Smooth or non-smooth? Smooth
3. Access to function/gradient evaluations? Function only
4. Noise/Stochastic oracle access? Noiseless

α-strictly convex and β-smooth

A continuously differentiable function is α-strictly convex if:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2} \|x - y\|^2 \quad \forall x, y \in K$$

It is β-smooth if:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2} \|x - y\|^2 \quad \forall x, y \in K$$

If the function is twice continuously differentiable, this is equivalent to its Hessian Hf having all eigenvalues in $[\alpha, \beta]$.

α-strictly convex and β-smooth: intuition

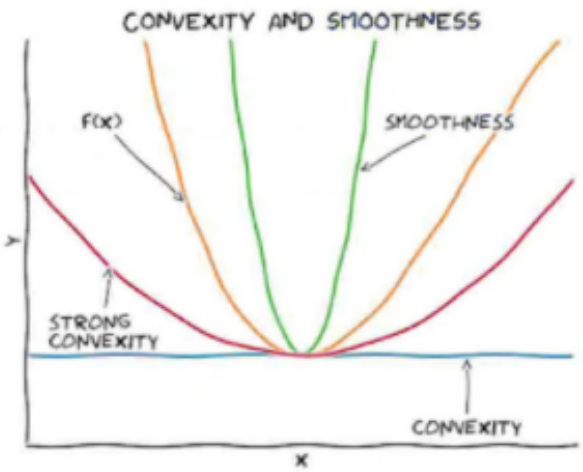

Figure: Intuition for α-strictly convex and β-smooth

Zeroth-order optimisation via GradientLess Descent

The algorithm

The paper introduces GradientLess Descent; it is based in the observation that for well-conditioned functions, taking a small step in a uniformly random direction has a high probability of reducing the objective function's value.

Algorithm 1: Gradientless Descent with Binary Search (GLD-Search)

Input: function: $f : \mathbb{R}^n \to \mathbb{R}$, $T \in \mathbb{Z}_+$: number of iterations, x_0: starting point,
$\mathcal{D}$: sampling distribution, R: maximum search radius, r: minimum search radius

Set $K = \log(R/r)$
for $t = 0, \ldots, T$ **do**
  Ball Sampling Trial i:
  for $k = 0, \ldots, K$ **do**
    Set $r_{i,k} = 2^{-k}R$.
    Sample $v_{i,k} \sim r_{i,k}\mathcal{D}$.
  end
  Update: $x_{t+1} = \arg\min_k \left\{ f(y) \middle| y = x_t, y = x_t + v_{i,k} \right\}$
end
return x_t

There is a clever way to pick the random sample which makes the algorithm faster.

Zeroth-order optimisation via GradientLess Descent

GLD beats ARS

Test setting: Let $H_{\alpha,\beta,n} \in M_n(\mathbb{R})$ be a diagonal matrix whose i-th diagonal is $d_i = \alpha + (\beta - \alpha)\frac{i-1}{n-1}$ and define $f_{\alpha,\beta,n} : \mathbb{R}^n \to \mathbb{R}$ by

$$f_{\alpha,\beta,n}(x) = \frac{1}{2} x^T H_{\alpha,\beta,n} x$$

So f is α-strongly convex and β-smooth.

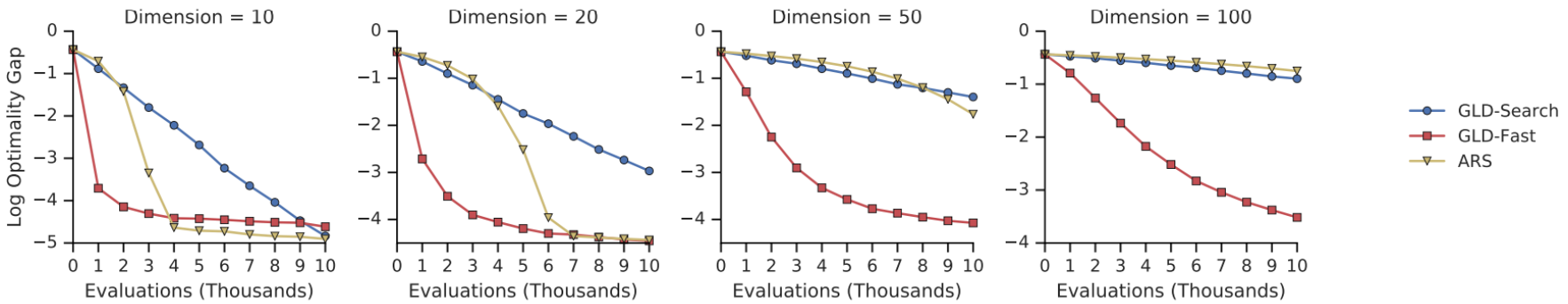

Figure: Plot of optimality gap $f(b_t) - f(x^*)$, where b_t is "best observed after t steps"

.

Zeroth-order optimisation via GradientLess Descent

Intersection of balls in high dimensions

Lemma

Let $B_1, B_2 \subseteq \mathbb{R}^n$ be balls of radii r_1, r_2. Let ℓ be the distance bweteen the centres. If $r_1 \in \left[\frac{\ell}{2\sqrt{n}}, \frac{\ell}{\sqrt{n}}\right]$ and $r_2 \geq \ell - \frac{\ell}{4n}$, then $\lambda(B_1 \cap B_2) \geq c_n \lambda(B_1)$, where $c_n \geq \frac{1}{4}$.

Intersection of balls in high dimensions

Lemma

Let $B_1, B_2 \subseteq \mathbb{R}^n$ be balls of radii r_1, r_2. Let ℓ be the distance bweteen the centres. If $r_1 \in \left[\frac{\ell}{2\sqrt{n}}, \frac{\ell}{\sqrt{n}}\right]$ and $r_2 \geq \ell - \frac{\ell}{4n}$, then $\lambda(B_1 \cap B_2) \geq c_n \lambda(B_1)$, where $c_n \geq \frac{1}{4}$.

Theorem

For any $x \in K$ such that $\frac{3}{5Q}\|x - x^*\| \in [C_1, C_2]$, we can find integers $0 \leq k_1, k_2 < \log \frac{C_2}{C_1}$ such that if $r = 2^{k_1} C_1$ or $r = 2^{-k_2} C_2$, then a sample y from the uniform distribution on $B_x = B\left(x, \frac{r}{\sqrt{n}}\right)$ satisfies

$$f(y) - f(x^*) \leq (f(x) - f(x^*))\left(1 - \frac{1}{5nQ}\right)$$

with probability at least $\frac{1}{4}$.

Bibliography I

[1] Daniel Golovin et al. "Gradientless descent: High-dimensional zeroth-order optimization". In: *arXiv preprint arXiv:1911.06317* (2019).

[2] Shengqiao Li. "Concise formulas for the area and volume of a hyperspherical cap". In: *Asian Journal of Mathematics and Statistics* 4.1 (2011), pp. 66–70.

[3] Yurii Nesterov and Vladimir Spokoiny. "Random gradient-free minimization of convex functions". In: *Foundations of Computational Mathematics* 17.2 (2017), pp. 527–566.

[4] R Tyrrell Rockafellar. *Convex analysis*. 28. Princeton university press, 1970.