A back-of-the-envelope treatment of FLOW2 on 2-dimensions

Iman Hosseini

March 19, 2021

1 FLOW2 on 2d

In "Frugal Optimization for Cost-related Hyperparameters" Wu et al. propose a randomized direct search method[1]. Abridged description: we are given a cost function $f(x) : \mathcal{R}^d -> \mathcal{R}$ and for a given step size and an initial point $\vec{x_0}$, we pick a direction $\vec{u}$, uniformly random from all possible directions, and compare $f(\vec{x_0})$ with $f(\vec{x_0} + \delta\vec{u})$ and $f(\vec{x_0} - \delta\vec{u})$ and if either of them is lower than $f(\vec{x_0})$ we pick $\vec{x_1}$ to be that new point (otherwise $\vec{x_1} = \vec{x_0}$) and we continue this procedure for a given number of steps (say, K).

Here we analyze the simple case of $d = 2$ to gain some intuition of the process, assuming a symmetric cost function only dependent on distance from a point. Denote with R_t the distance from the center at step t, we can illustrate a single step of this process as such: Here θ denotes the

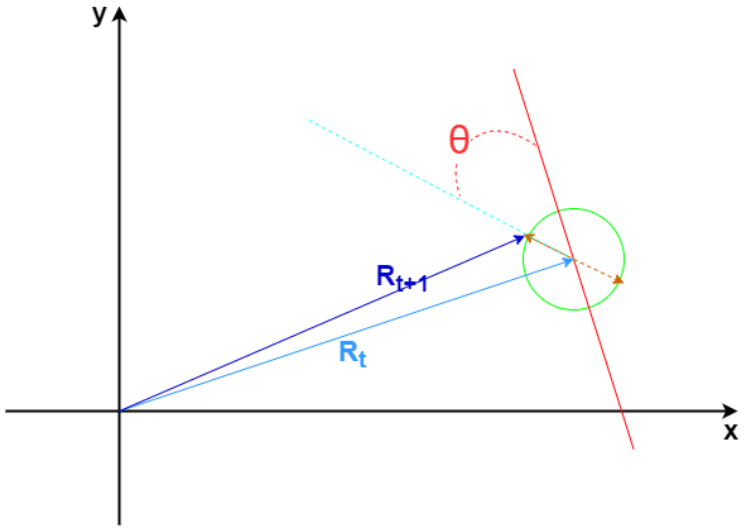

Figure 1: going from R_t to R_{t+1}

direction we pick, and notice how due to the symmetry of the two points under consideration $\vec{x} + \delta\vec{u}$, $\vec{x} - \delta\vec{u}$ being antipodal points on the 1-sphere we can assume that $\theta \in [0, \pi)$ coming from a uniform

[1] https://arxiv.org/abs/2005.01571

distribution. This symmetry, analogously persists at higher dimensions and is the gist of Lemma 1 in the paper. Also, (while $R > \delta$) that unless $\theta = 0$ which means that the direction we picked is perpendicular to $\vec{x_t}$, we would get a closer point to the center, in other words *we will almost surely* have $R_{t+1} < R_t$. Now to analyze the exact change in R we have -writing out cosine law-:

$$R_{t+1}^2 = R_t^2 + \delta^2 - 2\delta R_t sin(\theta) \tag{1}$$

For $\delta << R$ (where I intentionally drop index of R, meaning for all the R_ts under consideration) we can approximate the above relation up to first order in δ as:

$$R_{t+1} = R_t - \delta sin(\theta) \tag{2}$$

Taking the expected value (which is average over possible values for θ and I prefer to use the $<>$ notation for average in place of E) we get:

$$\begin{aligned} < R_{t+1} > &=< R_t - \delta sin(\theta) > \\ &=< R_t > - < \delta sin(\theta) > \\ &=< R_t > -\delta < sin(\theta) > \end{aligned} \tag{3}$$

The last part is easy to calculate:

$$\begin{aligned} < sin\theta > &= 1/\pi \int_{\theta=0}^{\theta=\pi} sin(\theta) d\theta = 2/pi \\ < R_{t+1} > &=< R_t > -\frac{2\delta}{\pi} \end{aligned} \tag{4}$$

As expected the $2/\pi \approx 0.6$ means that effectively, unlike the case of 1-dimension where we get closer to the center by δ, we don't approach the center as fast, in general this factor gets smaller for higher dimensions, and the expression for a general dimension d is derived in the appendix of Wu's paper. This shows how having more (hyper)parameters to tune means we should expect slower movement to the minimum of the loss function. Based on the above calculations, assuming for all t, it stays true that $R_t > \delta$ we would have after K steps:

$$< R_K > =< R_0 > -\frac{2K\delta}{\pi} \tag{5}$$

Here we only analyzed the behaviour of R_t *on average*, a more informative treatment would be to ask questions like: *if we want to be within ϵ with a probability of more than $1-\gamma$ how many steps should we take?* In this case, these questions become questions regarding the probability distribution of:

$$\Delta = \sum_{i=1}^{i=K} \delta sin(\theta_i) \tag{6}$$

Which is how much we moved towards the center, and it is clear that $0.0 \leq \Delta \leq K\delta$, so it makes sense to normalize it and speak of $\tilde{\Delta} = \Delta/K$. Now what we showed is that on average $\tilde{\Delta} = 2/\pi$ plotting the distribution of $\tilde{\Delta}$ for various K reveals (Fig. 2): Showing that *most probable* value is indeed $2/\pi$ but the variance grows high for smaller Ks. And notice that this is all for 2-dimension, in higher dimensions the variance would be even more.

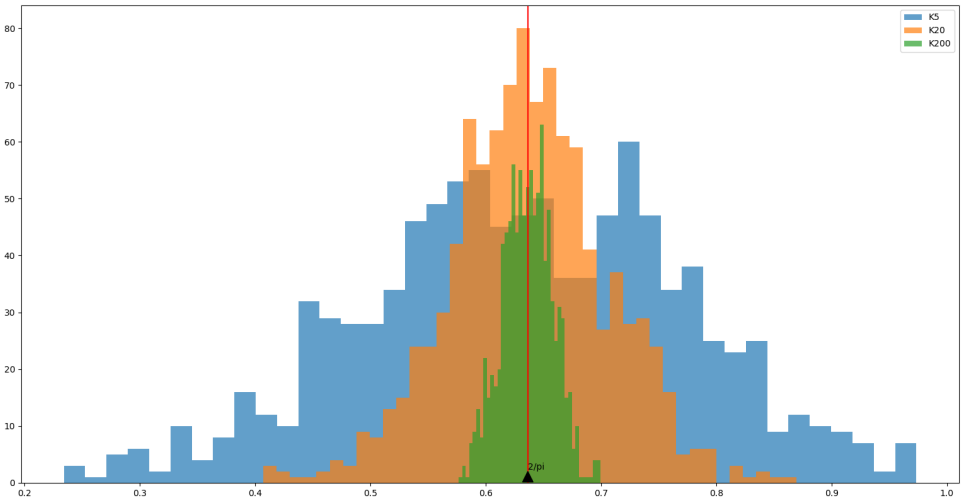

Figure 2: pdf of $\tilde{\Delta}_k$

2 Final Remarks

The idea of FLOW2 can be stated succinctly as: do a random walk over the search space, with the change that at each step also consider moving at the reverse direction, and only take a step if there is a step which makes $f(\vec{x_{t+1}}) < f(\vec{x_t})$. This construction can be readily generalized to other random processes, for example: the Lévy flight, where the steps are now changing. I predict that Lévy flight would lead to even better results, due to the *long steps* speeding up the approach towards the goal, and in fact previously Lévy flight has been applied to optimization problems.(one example I can point to is [2] but there is more)

[2] https://www.sciencedirect.com/science/article/pii/B978012405163800003X