# Gradient Free Convex Optimization

Pat Mellady

December 10, 2022

# Summary

# Why Study Derivative Free Methods

- Not all objective functions are differentiable
  - The authors present methods Gaussian smoothing techniques for minimization when $f$ is not differentiable.
- Derivative free methods are memory efficient
  - Don't need to store intermediate calculations
- More efficient use of programmer's time
  - Trade programmer's time for computing time

# Definitions, Theorems and Examples

## Our Problem

For a convex, differentiable function f with Lipschitz continuous gradient, we want to solve: $\min_{x \in \mathbb{R}} f(x)$

To perform gradient-free methods, we will use a numerical approximation of the directional derivative of $f$ in a random direction. We denote this by $g_\mu(x)$, as consistent with the author's notation, and define it below.

## Definition 1

$$g_\mu(x_k) = \frac{f(x_k + \mu u_k) - f(x_k)}{\mu} u_k$$

Where $u_k$ is a random vector whose entries are iid $N(0, 1)$.

## The Algorithm

Initialize: $x_0$, $\mu, N, k = 0$
Set the step size, h
While$(k < N)\{$
choose $u_k$ as described above
calculate $g_\mu(x_k)$
$x_{k+1} = x_k - hg_\mu(x_k)$
k=k+1
$\}$
return $x_{k+1}$

# Theoretical Results

## Theorem (Convergence and finding the step size)

Denote $\phi(k) = E_u(f(x_k))$ to be the expected value of $f(x_k)$ with respect to the random vectors, $u$. Let $f$ be convex and differentiable with Lipschitz continuous gradient. With a step size:

$$h = \frac{1}{4(n+4)L_1}$$

where n is the dimension of the problem and $L_1$ is the Lipschitz constant for the gradient. Then, for any iteration limit, N, we have:

$$\frac{1}{N+1}\sum_{k=0}^{N}(\phi(k) - f(x^*)) \leq \frac{4(n+4)L_1||x_0 - x^*||^2}{N+1} + \frac{9\mu^2(n+4)^2 L_1}{25}$$

The above tells us that, with a clever selection of the step size, we can place probabilistic convergence bounds on the average error over the whole history of the algorithm.

# Theoretical Results Continued

## How can we optimize the choice of $\mu$

In order to obtain accuracy within $\epsilon$ of the actual minimizer, we can cleverly choose

- $\mu = \frac{5}{3(n+4)}\sqrt{\frac{\epsilon}{2L}}$

Where $\epsilon$ is the desired accuracy of our minimizer

## How does this theorem characterize convergence?

As stated in the definition of the algorithm, we have choice $\mu$ and N

- We can make $\mu$ small by requiring high accuracy (small $\epsilon$)
- We can make N large

This allows us to make the error bound as small as we desire

# Drawbacks

- Not truly derivative free
  - While we do not need to compute the gradient, in this algorithm, we are approximating the directional derivative in a clever way
- Even though the method is gradient free, f must have Lipschitz gradient
  - Finding the Lipschitz constant to be used in the calculation for step size and $\mu$ is not necessarily easy
- Not easily applied to constrained optimization
  - Projecting each iteration onto the feasible set and finding projection mappings is not, in general, easy
- Slower and Weaker Convergence
  - The algorithm is slower to converge and probabilistic convergence is weaker than the deterministic convergence.

# Example

To showcase the method, we will use the random gradient free method and test it against the method of gradient descent. The function we will minimize is the Booth function

$$f(x_1, x_2) = (x_1 + 2x_2 - 7)^2 + (2x_1 + x_2 - 5)^2$$

Which has a minimum of $f(1, 3) = 0$

# References

Nesterov, Y., Spokoiny, V. *Random Gradient-Free Minimization of Convex Functions. Found Comput Math 17, 527–566 (2017).* https://doi.org/10.1007/s10208-015-9296-2