A weak tail-bound probabilistic condition for function estimation in stochastic derivative-free optimization (with improved sample sizing)

Luis Nunes Vicente

joint work with Francesco Rinaldi & Damiano Zeffiro

12th US–Mexico Workshop on Optimization and its Applications

Steve’s “60th” Birthday

January 11, 2023
1 Introduction

2 The tail bound probabilistic condition & sample sizing

3 Numerical experiments

4 Let’s take a break

5 A simple stochastic direct-search scheme

6 A simple stochastic trust-region scheme

7 Conclusions and extensions
Problem formulation

\[
\min_{x \in \mathbb{R}^n} f(x)
\]

where \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) is

- locally Lipschitz continuous
- possibly non-smooth and with \( \inf f = f^* \)
- given by a stochastic oracle

\[
F(x, \xi) \approx f(x)
\]

with oracle given by sampling over \( \xi \).
Some notation

- Probability space \((\mathbb{P}, \Omega, \mathcal{F})\)

- \(w\) outcome of the sample space \(\Omega\)

- Our algorithms generate random processes:
  - \(g_k\) direction realization (shorthand for\(G_k(w)\))
  - \(\delta_k\) stepsize realization (shorthand for \(\Delta_k(w)\))
  - \(f_k\) estimate realization for \(f(x_k)\) (shorthand for \(F_k(w)\))
  - same for \(f^g_k \sim f(x_k + \delta_k g_k)\)

- \(\mathcal{F}_{k-1}\) is the \(\sigma\)–algebra of events up to the choice of \(g_k\)

- The acceptance criterion is \(f_k - f^g_k \geq \theta \delta^q_k\), for \(\theta > 0, q > 1\)
Tail-bound probabilistic condition

Assumption (Tail bound)

For some $\varepsilon_q > 0$ (independent of $k$):

$$\mathbb{P} \left( \left| F_k - F_k^g - (f(X_k) - f(X_k + \Delta_k G_k)) \right| \geq \alpha \Delta_k^q |\mathcal{F}_{k-1}\right) \leq \frac{\varepsilon_q}{\alpha^{q/(q-1)}}$$

a.s. for every $\alpha > 0$.

- power law tail bound on error with exponent $q/(q-1)$
**Tail-bound probabilistic condition**

**Assumption (Tail bound)**

*For some $\varepsilon_q > 0$ (independent of $k$):*

$$\mathbb{P} \left( |F_k - F^g_k - (f(X_k) - f(X_k + \Delta_k G_k))| \geq \alpha \Delta_k^q |F_{k-1}| \right) \leq \frac{\varepsilon_q}{\alpha q/(q-1)}$$

a.s. for every $\alpha > 0$.

- power law tail bound on error with exponent $q/(q - 1)$
- satisfied, since if $r$-moment of noise is finite ($r \geq 2$), then:

$$\mathbb{E}(|A_k|^r) \leq C_r p_k^{-\frac{r}{2}}$$

when $A_k = F_k - F^g_k - (f(X_k) - f(X_k + \Delta_k G_k))$ considers averaging $p_k$ i.i.d. samples in $F_k, F^g_k$ (and that estimator is unbiased)
Assumption (Bounded moment)

For some $r > 1$, \[ \mathbb{E}_\xi [ |F(x, \xi) - f(x)|^r ] \leq M_r < +\infty \]
Sample bound for bounded moment – (i)

Assumption (Bounded moment)

For some \( r > 1 \), \( \mathbb{E}_\xi[|F(x, \xi) - f(x)|^r] \leq M_r < +\infty \)

Theorem

Assume the estimator for \( A_k \) is unbiased (true if \( f(x) = \mathbb{E}_\xi[F(x, \xi)] \)).

When \( r = r(q) = \frac{q}{q-1} \), \( q \in (1, 2] \), the tail bound can be satisfied by averaging

\[
O\left(\Delta_k^{-2q}\right) \quad i.i.d. \ samples
\]

- for \( q = 1.5 \) \( (r = 3) \) only \( O(\Delta_k^{-3}) \) samples needed
- for \( q = 2 \) \( (r = 2) \) the known bound is \( O(\Delta_k^{-4}) \)
Use of $r$-th moment and $q,r$ being conjugates:

$$P(|A| \geq \alpha \Delta^{\frac{r}{r-1}})$$
Use of $r$-th moment and $q, r$ being conjugates:

$$
P(|A| \geq \alpha \Delta^{\frac{r}{r-1}}) = P(|A|^r \geq \alpha^r \Delta^{\frac{r^2}{r-1}})
$$
Use of $r$-th moment and $q,r$ being conjugates:

\[
\mathbb{P}(|A| \geq \alpha \Delta^{\frac{r}{r-1}}) = \mathbb{P}(|A|^r \geq \alpha^r \Delta^{\frac{r^2}{r-1}}) \\
\leq \frac{\mathbb{E}[|A|^r]}{\alpha^r \Delta^{r^2/(r-1)}}
\]
Sample bound for bounded moment – (ii)

Use of $r$-th moment and $q, r$ being conjugates:

$$\mathbb{P}( |A| \geq \alpha \Delta^{r/(r-1)} ) = \mathbb{P}( |A|^r \geq \alpha^r \Delta^{r^2/(r-1)} )$$

$$\leq \frac{\mathbb{E}[|A|^r]}{\alpha^r \Delta^{r^2/(r-1)}} \leq \frac{2^r C_r M_r p^{-r/2}}{\alpha^r \Delta^{r^2/(r-1)}}$$
Sample bound for bounded moment – (ii)

Use of $r$-th moment and $q, r$ being conjugates:

$$
\mathbb{P}(|A| \geq \alpha \Delta^{r/(r-1)}) = \mathbb{P}(|A|^r \geq \alpha^r \Delta^{r^2/(r-1)}) \\
\leq \frac{\mathbb{E}[|A|^r]}{\alpha^r \Delta^{r^2/(r-1)}} \leq \frac{2^r C_r M_r p^{-\frac{r}{2}}}{\alpha^r \Delta^{r^2/(r-1)}} = \frac{\varepsilon_q}{\alpha^r}
$$

for $p = O(\Delta^{\frac{-2r}{r-1}}) = O(\Delta^{-2q})$. 


Suppose we have access to the random number generator (we can fix $\xi$ and sample $F(\cdot, \xi)$), and the errors are correlated in the form:

**Assumption (Correlated error)**

Let $\bar{F}(x, \xi) = F(x, \xi) - f(x)$. For some $r > 1$:

$$\mathbb{E}_\xi[|\bar{F}(x, \xi) - \bar{F}(y, \xi)|^r] \leq D_r \|x - y\|^r$$
Correlated errors

Suppose we have access to the random number generator (we can fix $\xi$ and sample $F(\cdot, \xi)$), and the errors are correlated in the form:

**Assumption (Correlated error)**

Let $\bar{F}(x, \xi) = F(x, \xi) - f(x)$. For some $r > 1$:

$$
E_\xi[|\bar{F}(x, \xi) - \bar{F}(y, \xi)|^r] \leq D_r \|x - y\|^r
$$

• ensured, for every $r$, when $F(x, \xi)$ is a Gaussian process with exponentiated quadratic kernel $K(x, y) = \sigma^2 \exp\left(-\frac{\|x-y\|^2}{2l^2}\right)$

  in which case $\text{Var}_\xi[F(x, \xi)]$ is constant and

  $$
  \text{Cov}_\xi(F(x, \xi), F(y, \xi)) \geq O \left(1 - \|x - y\|^2\right)
  $$
Sample bound for correlated errors

Theorem

Assume the estimator for $A_k$ is unbiased (true if $f(x) = \mathbb{E}_\xi[F(x, \xi)]$).

When $r = \frac{q}{q-1}$, $q \in (1, 2]$, the tail bound can be satisfied by averaging:

$$O(\Delta_k^{2-2q}) \quad \text{i.i.d. samples}$$

- for $q = 1.5$ ($r = 3$) only $O(\Delta_k^{-1})$ samples needed
- for $q = 2$ ($r = 2$) one gets $O(\Delta_k^{-2})$
tested the direct-search algorithm for \( q \in \{1.5, 2\} \), for which \( r(q) \in \{3, 2\} \)

algorithms tested on a set of 96 well known non-smooth problems

added Gaussian noise \( N(0, 10^{-2}) \) in the general case, \( N(0, \delta_k 10^{-2}) \) in the correlated one

for the moment bound case, number of samples was: \( \lceil \delta_k^{-4} \rceil \) \( (q = 2) \) and \( \lceil \delta_k^{-3} \rceil \) \( (q = 1.5) \)

for the correlated errors case, number of samples was: \( \lceil \delta_k^{-2} \rceil \) \( (q = 2) \) and \( \lceil \delta_k^{-1} \rceil \) \( (q = 1.5) \)

data and performance profiles
**Figure:** From left to right, data and performance profiles. From top to bottom, tolerance $10^{-2}$ and $10^{-4}$
Figure: From left to right, data and performance profiles. From top to bottom, tolerance $10^{-2}$ and $10^{-4}$
Is there an optimal $q$ in $(1,2]$?
When \( F(x, \varepsilon) \sim f(x) \sim N(0, \sigma) \), the tail bound condition is satisfied using

\[
p = B(q) := \left| \frac{4\sigma^2 M_{r(q)}^{2/r(q)}}{\varepsilon^{2/r(q)}} \Delta^{-2q} \right|
\]

where \( r(q) = \frac{q}{q-1} \) and \( M_{r(q)} \) is the \( r(q) \)-th moment of a standard normal distribution.

The continuous version of \( B(q) \) has always a minimum in \((1, 2]\).
Comparison with other assumptions – 1

\(k_f\)-variance conditions [Audet et al., 2021]

\[
\mathbb{E}[|F_k^g - f(X_k + \Delta_k G_k)|^2 | \mathcal{F}_{k-1}] \leq k_f^2 \Delta_k^4
\]

\[
\mathbb{E}[|F_k - f(X_k)|^2 | \mathcal{F}_{k-1}] \leq k_f^2 \Delta_k^4
\]

**Proposition**

*Then tail bound condition is satisfied for \(\varepsilon_q = 4k_f^2\) and \(q = 2\).*

- follows from Markov’s inequality
Comparison with other assumptions – 2

\(\beta\)-probabilistically accurate function estimate [Chen et al. 2018]

\[
\mathbb{P}(\{|F_k - f(X_k)| \leq \tau_f \Delta_k^2\} \cap \{|F^g_k - f(X_k + \Delta_k G_k)| \leq \tau_f \Delta_k^2\} \mid \mathcal{F}_{k-1}) \geq \beta
\]

Proposition

If satisfied for all \(\beta\) in a chosen interval (and \(\tau_f\) depending on \(\beta\) and accuracy parameter \(\varepsilon\)), then tail bound is satisfied with \(\varepsilon_q\) depending on \(\varepsilon\).

follows from the inclusion

\[
\{|F_k - F^g_k - (f(X_k) - f(X_k + \Delta_k G_k))| < \alpha \Delta_k^2\}
\]

\[
\supset \{|F_k - f(X_k)| \leq \tau_f \Delta_k^2\} \cap \{|F^g_k - f(X_k + \Delta_k G_k)| \leq \tau_f \Delta_k^2\}
\]

for any \(\tau_f < \frac{\alpha}{2}\).
Let’s take a break...

Apologies for all vegans and vegetarians...

I am also celebrating the 25th anniversary of Steve’s 2-week visit to Portugal...

Here is a quiz for Steve... Let’s test his memory in real time. :-}
What are we eating here?
Algorithm Stochastic direct search

1: **Initialization.** Choose a point $x_0$, $\delta_0$, $\theta > 0$, $\tau \in (0, 1)$, $\bar{\tau} \in [1, 1 + \tau]$.
2: **For** $k = 0, 1 \ldots$
3: Select a direction $g_k$ in the unitary sphere.
4: Compute estimates $f_k$ and $f^g_k$ for $f$ in $x_k$ and $x_k + \delta_k g_k$.
5: **If** $f_k - f^g_k \geq \theta \delta^g_k$, **Then** set $x_{k+1} = x_k + \delta_k g_k$, $\delta_{k+1} = \bar{\tau} \delta_k$.
6: **Else** set $x_{k+1} = x_k$, $\delta_{k+1} = (1 - \tau) \delta_k$.
7: **End if**
8: **End for**
Bad successful step

Figure: A bad successful step
Tail-bound probabilistic condition (again)

Assumption (Tail bound)

For some $\varepsilon_q > 0$ (independent of $k$):

$$
P \left( |F_k - F_k^g - (f(X_k) - f(X_k + \Delta_k G_k))| \geq \alpha \Delta_k^q \big| \mathcal{F}_{k-1} \right) \leq \frac{\varepsilon_q}{\alpha^q/(q-1)}
$$

a.s. for every $\alpha > 0$. 
Lemma

Under the tail bound condition, if $\theta > \theta^{ds}(q, \tau, \varepsilon_q)$, then a.s.

$$\sum \Delta^q_k < +\infty$$

- Let $\Phi_k = f(X_k) - f^* + C_1 \Delta^q_k$
- The lemma follows from Robbins-Siegmund once we get to

$$\mathbb{E}[\Phi_k - \Phi_{k+1}|\mathcal{F}_{k-1}] \geq C_2 \Delta^q_k$$

- For a certain $\rho_k$, the above LHS is $\geq$ than

$$\left(C_3 - \rho_k \text{ (in tail bound with } \alpha = \rho_k) \Delta^q_k \right) \leq C_4(1/\rho_k)$$
For some $\varepsilon_q > 0$ (independent of $k$):

$$\mathbb{P} \left( |F_k - F^g_k - (f(X_k) - f(X_k + \Delta_k G_k))| \geq \alpha \Delta_k^q \mid \mathcal{F}_{k-1} \right) \leq \frac{\varepsilon_q}{\alpha^q/(q-1)}$$

a.s. for every $\alpha > 0$. 

Assumption (Tail bound)
Let $K$ be the set of indices of unsuccessful iterations. Then under the tail bound assumption and $\theta > \theta_{ds}$ we have a.s.

$$\liminf_{k \in K, k \to \infty} \frac{f(X_k + \Delta_k G_k) - f(X_k)}{\Delta_k} \geq 0$$

- need to prove $|F_k - F^g_k - (f(X_k) - f(X_k + \Delta_k G_k))|/\Delta_k \to 0$
- apply the tail bound assumption with $\alpha = \frac{\Delta_k^{1-q}}{m}$

$$\mathbb{P}(|F_k - F^g_k - (f(X_k) - f(X_k + \Delta_k G_k))| \geq \frac{\Delta_k}{m} \mid \mathcal{F}_{k-1}) \leq m^{r(q)} \Delta_k^{q} \varepsilon_q$$

- conclusion from Borel-Cantelli’s First Lemma for every $m$
Convergence to Clarke stationary points

Theorem

Let the tail bound assumption hold, \( \theta > \theta^{ds} \), and \( f \) Lipschitz continuous around any limit point.

If \( L \subseteq K \) is such that \( \{G_k\}_{k \in L} \) is dense in the unit sphere and

\[
\lim_{k \in L, k \to \infty} X_k = X^*
\]

then \( X^* \) is Clarke stationary (a.s.).

- follows from last lemma and \( \limsup \geq \liminf \) (and \( \Delta_k \to 0 \))
A simple stochastic trust-region scheme

**Algorithm** Stochastic DFO Trust-Region Algorithm

1: **Initialization.** Select $x_0 \in \mathbb{R}^n$, $\theta > 0$, $\tau \in (0, 1)$, $\bar{\tau} \in [1, 1 + \tau]$, $\delta_0 > 0$, $q > 1$.
2: **For** $k = 0, 1 \ldots$
3: Select a direction $g_k \neq 0$ and build a symmetric matrix $B_k$.
4: Compute $s_k \in \arg\min \|s\| \leq \delta_k g_k^\top s + \frac{1}{2} s^\top B_k s$.
5: Compute estimates $f_k \simeq f(x_k)$ and $f_s \simeq f(x_k + s_k)$.
6: **If** $\frac{f_k - f_s}{\theta \|s_k\|^q} \geq 1$
   **Then** set $x_{k+1} = x_k + s_k$, $\delta_{k+1} = \bar{\tau} \delta_k$.
7: **Else** set $x_{k+1} = x_k$, $\delta_{k+1} = (1 - \tau) \delta_k$.
8: **End For**
Assumption (Trust-region tail bound)

For some $\varepsilon_q > 0$ (independent of $k$):

$$\mathbb{P} \left( |F_k - F^g_k - (f(X_k) - f(X_k + S_k))| \geq \alpha \|S_k\|^q |\mathcal{F}_{k-1} \right) \leq \frac{\varepsilon_q}{\alpha^{q/(q-1)}}$$

a.s. every $\alpha > 0$.

- $S_k, \|S_k\|, F^s_k$ replace $\Delta_k G_k, \Delta_k, F^g_k$
- same improved sampling bounds of direct-search case
Under the tail bound condition

\[
\sum \|S_k\|^q < +\infty
\]

for a different lower bound \( \theta > \theta^{tr}(q, \tau, \varepsilon_q, \rho) \).

Assumption (Hessian bound 1)

There exists \( \rho \in (0, 1] \) such that, for every \( k \),

\[
\|B_k\| \leq \frac{1}{\rho} \frac{\|G_k\|}{\Delta_k}
\]

- when \( \|G_k\| = 1 \), Hessian is “unbounded” by \( 1/\Delta_k \)
- it implies \( \|S_k\| \geq \rho \Delta_k \), which then gives \( \sum \Delta_k^q < +\infty \)
Assumption (Hessian bound 2)

There exists a sequence \( \{a_k\} \downarrow 0 \) and such that, for every \( k \),

\[
\|B_k\| \leq a_k \frac{\|G_k\|}{\Delta_k}
\]

Lemma (asymptotic alignment)

If \( S_k \) solves the trust-region subproblem,

\[
\lim_{k \to \infty} \frac{G_k}{\|G_k\|} + \frac{S_k}{\|S_k\|} = 0
\]

a.s. (it holds for every realization, actually).

- for \( k \) large, \( S_k \) becomes aligned with \(-G_k\)
Let the tail bound assumption hold, $\theta > \theta^{tr}$, $f$ Lipschitz continuous around any limit point, and Hessian bound 2.

If $L \subset K$ is such that $\{G_k\}_{k \in L}$ is dense in the unit sphere and

$$\lim_{k \in L, k \to \infty} X_k = X^*$$

then $X^*$ is Clarke stationary (a.s.).

- corollary of analogous DS result for $\left\{ \frac{S_k}{\|S_k\|} \right\}$ + asymptotic alignment
Conclusions and extensions

Conclusions
- introduced a tail bound condition tailored to acceptance criterion
- proved improved bounds on the corresponding number of samples
- proved convergence of a direct-search and a trust-region schemes

Extensions
- more general random trust-region models (e.g. piecewise linear)
- composition of smooth function with known non-smooth function
- numerical experiments for trust-region method