Direct search based on probabilistic descent in reduced subspaces

Clément W. Royer (Université Paris Dauphine-PSL)

Joint work with Lindon Roberts (Univ. of Sydney)

SIAM OP - June 3, 2023
The paper

*Direct search based on probabilistic descent in reduced spaces*

About this work

The paper

Direct search based on probabilistic descent in reduced spaces


Since the paper came out...

We got one round of reviews;
We had three kids (1+2);
We got a lot of feedback and people even moved this forward!
Thank you all and huge thanks to Lindon!
About this work

The paper

*Direct search based on probabilistic descent in reduced spaces*


Since the paper came out...

- We got one round of reviews;
About this work

The paper

*Direct search based on probabilistic descent in reduced spaces*


Since the paper came out...

- We got one round of reviews;
- We had three kids (1+2);
The paper

*Direct search based on probabilistic descent in reduced spaces*


Since the paper came out...

- We got one round of reviews;
- We had three kids (1+2);
- We got a lot of feedback and people even moved this forward!
The paper

*Direct search based on probabilistic descent in reduced spaces*


Since the paper came out...

- We got one round of reviews;
- We had three kids (1+2);
- We got a lot of feedback and people even moved this forward!

Thank you all and huge thanks to Lindon!
1. Direct search
2. Probabilistic descent
3. In reduced subspaces
4. It works
Direct search

Probabilistic descent

In reduced subspaces

It works
**Introductory assumptions and definitions**

\[
\text{minimize}_{x \in \mathbb{R}^n} \ f(x).
\]

**Assumptions**
- \( f \) bounded below;
- \( f \) continuously differentiable (nonconvex).

**Blackbox/Derivative-free setup**
- Derivatives unavailable for algorithmic use.
- Only access to values of \( f \).
My goal as a derivative-free/blackbox optimizer

Develop algorithms with controlled

- Number of calls to $f$;
- Dependency on $n$. 

Complexity bound

Given $\epsilon \in (0, 1)$ and, bound the number of function evaluations needed by a method to reach $x$ such that $\|\nabla f(x)\| \leq \epsilon$, deterministically or in expectation/probability.
Complexity in blackbox optimization

My goal as a derivative-free/blackbox optimizer

Develop algorithms with controlled
- Number of calls to $f$;
- Dependency on $n$.

Complexity bound

Given $\epsilon \in (0, 1)$ and, bound the number of function evaluations needed by a method to reach $x$ such that

$$\|\nabla f(x)\| \leq \epsilon,$$

deterministically or in expectation/probability.
My goal as a derivative-free/blackbox optimizer

Develop algorithms with controlled

- Number of calls to \( f \);
- Dependency on \( n \).

Complexity bound

Given \( \varepsilon \in (0, 1) \) and, bound the number of function evaluations needed by a method to reach \( \mathbf{x} \) such that

\[
\| \nabla f(\mathbf{x}) \| \leq \varepsilon,
\]

deterministically or in expectation/probability.

**Focus:** dependency w.r.t. \( n \).
A (simplified) direct-search framework

Inputs: \( x_0 \in \mathbb{R}^n \) \( 0 < \gamma_{\text{dec}} < 1 \leq \gamma_{\text{inc}}, \alpha_0 > 0 \).

Iteration \( k \): Given \( (x_k, \alpha_k) \),

- Choose a set \( D_k \subset \mathbb{R}^n \) of \( m \) vectors.
- If \( \exists \, d_k \in D_k \) such that

\[
f(x_k + \alpha_k d_k) < f(x_k) - \alpha_k^2 \|d_k\|^2
\]

set \( x_{k+1} := x_k + \alpha_k d_k, \alpha_{k+1} := \gamma_{\text{inc}} \alpha_k \).

- Otherwise, set \( x_{k+1} := x_k, \alpha_{k+1} := \gamma_{\text{dec}} \alpha_k \).
A (simplified) direct-search framework

**Inputs:** $x_0 \in \mathbb{R}^n$ $0 < \gamma_{\text{dec}} < 1 \leq \gamma_{\text{inc}}$, $\alpha_0 > 0$.

**Iteration $k$:** Given $(x_k, \alpha_k)$,

- Choose a set $D_k \subset \mathbb{R}^n$ of $m$ vectors.
- If $\exists d_k \in D_k$ such that
  \[
  f(x_k + \alpha_k d_k) < f(x_k) - \alpha_k^2 \|d_k\|^2
  \]
  set $x_{k+1} := x_k + \alpha_k d_k$, $\alpha_{k+1} := \gamma_{\text{inc}} \alpha_k$.
- Otherwise, set $x_{k+1} := x_k$, $\alpha_{k+1} := \gamma_{\text{dec}} \alpha_k$. 

Which vectors should we use?

C. W. Royer

Direct search in reduced subspaces

SIOP 23 7
A (simplified) direct-search framework

**Inputs:** \( x_0 \in \mathbb{R}^n \) \( 0 < \gamma_{\text{dec}} < 1 \leq \gamma_{\text{inc}}, \alpha_0 > 0 \).

**Iteration** \( k \): Given \((x_k, \alpha_k)\),

- Choose a set \( D_k \subset \mathbb{R}^n \) of \( m \) vectors.
- If \( \exists d_k \in D_k \) such that
  
  \[
  f(x_k + \alpha_k d_k) < f(x_k) - \alpha_k^2 \|d_k\|^2
  \]

  set \( x_{k+1} := x_k + \alpha_k d_k, \alpha_{k+1} := \gamma_{\text{inc}} \alpha_k \).
- Otherwise, set \( x_{k+1} := x_k, \alpha_{k+1} := \gamma_{\text{dec}} \alpha_k \).

**Which vectors should we use?**
Choosing $\mathcal{D}_k$

### A measure of set quality

The set $\mathcal{D}_k$ is called $\kappa$-descent for $f$ at $x_k$ if

$$
\max_{d \in \mathcal{D}_k} -\frac{d^T \nabla f(x_k)}{\|d\| \|\nabla f(x_k)\|} \geq \kappa \in (0, 1].
$$
Choosing $\mathcal{D}_k$

A measure of set quality

The set $\mathcal{D}_k$ is called $\kappa$-descent for $f$ at $x_k$ if

$$\max_{d \in \mathcal{D}_k} \frac{-d^T \nabla f(x_k)}{\|d\| \|\nabla f(x_k)\|} \geq \kappa \in (0, 1].$$

- Guaranteed when $\mathcal{D}_k$ is a Positive Spanning Set (PSS);
- $\mathcal{D}_k$ PSS $\Rightarrow |\mathcal{D}_k| \geq n + 1$;
- Ex) $\mathcal{D}_\oplus := \{e_1, \ldots, e_n, -e_1, \ldots, -e_n\}$ is always $\frac{1}{\sqrt{n}}$-descent.
Assumption: For every $k$, $D_k$ is $\kappa$-descent and contains $m$ unit directions.

Theorem

Let $\epsilon \in (0, 1)$ and $N_\epsilon$ be the number of function evaluations needed to reach $x_k$ such that $\|\nabla f(x_k)\| \leq \epsilon$. Then,

$$N_\epsilon \leq O \left( m \kappa^{-2} \epsilon^{-2} \right).$$
Worst-case complexity in deterministic direct search

**Assumption:** For every $k$, $D_k$ is $\kappa$-descent and contains $m$ unit directions.

**Theorem**

Let $\epsilon \in (0, 1)$ and $N_\epsilon$ be the number of function evaluations needed to reach $x_k$ such that $\|\nabla f(x_k)\| \leq \epsilon$. Then,

$$N_\epsilon \leq O\left( m \kappa^{-2} \epsilon^{-2} \right).$$

- Unit norm can be replaced by bounded norm.
- Choosing $D_k = D_\oplus$, one has $\kappa = \frac{1}{\sqrt{n}}$, $m = 2n$, and the bound becomes

  $$N_\epsilon \leq O\left( n^2 \epsilon^{-2} \right).$$

  $\Rightarrow$ **Optimal** in the power of $n$ for deterministic direct-search algorithms.
Roadmap

1. Direct search
2. Probabilistic descent
3. In reduced subspaces
4. It works
### Deterministic descent

The set $\mathcal{D}_k$ is $\kappa$-descent for $(f, x_k)$ if

$$\max_{d \in \mathcal{D}_k} \frac{-\nabla f(x_k)^T d}{\|\nabla f(x_k)\| \|d\|} \geq \kappa \in (0, 1].$$
A probabilistic property

Deterministic descent

The set $\mathcal{D}_k$ is $\kappa$-descent for $(f, x_k)$ if

$$\max_{d \in D_k} \frac{-\nabla f(x_k)^\top d}{\|\nabla f(x_k)\| \|d\|} \geq \kappa \in (0, 1].$$

Probabilistic descent

The sequence $\{\mathcal{D}_k\}$ is said to be $(p, \kappa)$-descent if:

$$\mathbb{P}(\mathcal{D}_0 \text{ $\kappa$-descent}) \geq p$$

$$\forall k \geq 1, \quad \mathbb{P}(\mathcal{D}_k \text{ $\kappa$-descent} \mid \mathcal{D}_0, \ldots, \mathcal{D}_{k-1}) \geq p,$$
Assumptions:
- \( \{ D_k \} (p, \kappa)\)-descent, \( p > p_0 = p_0(\gamma_{\text{inc}}, \gamma_{\text{dec}}) \).
- \( D_k \) contains \( m \) unit vectors.

Probabilistic worst-case complexity (Gratton et al, ’15)

Let \( \epsilon \in (0, 1) \) and \( N_\epsilon \) the number of function evaluations needed to have \( \| \nabla f(x_k) \| \leq \epsilon \). Then

\[
P \left( N_\epsilon \leq \mathcal{O} \left( \frac{m \kappa^{-2} \epsilon^{-2}}{p - p_0} \right) \right) \geq 1 - \exp \left( -\mathcal{O} \left( \frac{p - p_0}{p} (\kappa \epsilon)^{-2} \right) \right).
\]
Complexity results

Assumptions:
- \( \{D_k\} \) \((p, \kappa)\)-descent, \( p > p_0 = p_0(\gamma_{\text{inc}}, \gamma_{\text{dec}}) \).
- \( D_k \) contains \( m \) unit vectors.

Probabilistic worst-case complexity (Gratton et al, ’15)

Let \( \epsilon \in (0, 1) \) and \( N_\epsilon \) the number of function evaluations needed to have \( \|\nabla f(x_k)\| \leq \epsilon \). Then
\[
\mathbb{P}\left( N_\epsilon \leq \mathcal{O}\left( \frac{m \kappa^{-2} \epsilon^{-2}}{p - p_0} \right) \right) \geq 1 - \exp\left( -\mathcal{O}\left( \frac{p - p_0}{p} (\kappa \epsilon)^{-2} \right) \right).
\]

Expected evaluation complexity
\[
\mathbb{E}[N_\epsilon] \leq \mathcal{O}\left( \frac{m \kappa^{-2} \epsilon^{-2}}{p - p_0} \right) + \mathcal{O}(m).
\]
Using a vector uniformly distributed over the unit sphere and its negative.

- Defines a $(p, \mathcal{O}(1/\sqrt{n}))$-descent sequence, $p > p_0$.
- Uses a one-dimensional subspace.

Complexity bound

Deterministic: $m = \mathcal{O}(n) \Rightarrow \mathcal{O}(n^{2\epsilon - 2})$.

Probabilistic: $m = \mathcal{O}(1) \Rightarrow \mathcal{O}(n^{\epsilon - 2})$.

⇒ Factor $n$ improvement at the iteration level.
Using a vector uniformly distributed over the unit sphere and its negative.

- Defines a $(p, \mathcal{O}(1/\sqrt{n}))$-descent sequence, $p > p_0$.
- Uses a one-dimensional subspace.

Complexity bound

- Deterministic: $m = \mathcal{O}(n) \Rightarrow \mathcal{O}(n^2 \epsilon^{-2})$.
- Probabilistic $m = \mathcal{O}(1) \Rightarrow \mathcal{O}(n \epsilon^{-2})$.
  \[ \Rightarrow \text{Factor } n \text{ improvement at the iteration level.} \]
Gaussian smoothing approach: Draw $u_k \sim \mathcal{N}(0, I)$ and use

$$\frac{f(x + \alpha u) - f(x)}{\alpha} u \quad \text{or} \quad \frac{f(x + \alpha u) - f(x - \alpha u)}{\alpha} u.$$ 

Random gradient-free method (Nesterov and Spokoiny 2017), Stochastic three-point method (Bergou et al, 2020).

- Also achieve $O(n\epsilon^{-2})$ bound.
- Use one-dimensional subspace based on Gaussian vectors.
- Use fixed or decreasing stepsizes.
Gaussian smoothing approach: Draw $u_k \sim \mathcal{N}(0, I)$ and use

$$\frac{f(x + \alpha u) - f(x)}{\alpha} u$$ or $$\frac{f(x + \alpha u) - f(x - \alpha u)}{\alpha} u.$$ 

Random gradient-free method (Nesterov and Spokoiny 2017), Stochastic three-point method (Bergou et al, 2020).

- Also achieve $O(n\epsilon^{-2})$ bound.
- Use one-dimensional subspace based on Gaussian vectors.
- Use fixed or decreasing stepsizes.

Our questions

- Gaussian directions are not always bounded
  \( \Rightarrow \) Can we extend the probabilistic analysis?
- Can we do better than $O(n)$?
- What should we do about the stepsizes?
1 Direct search
2 Probabilistic descent
3 In reduced subspaces
4 It works
Derivative-free methods and subspaces

Model based

- Build a model of the objective;
- Recent interest in building models over random subspaces (Cartis and Roberts ’23, Dzahini and Wild ’22a).
Derivative-free methods and subspaces

Model based

- Build a model of the objective;
- Recent interest in building models over random subspaces (Cartis and Roberts ’23, Dzahini and Wild ’22a).

Direct search

- Sample along appropriate directions;
- Done before: Use random directions in one-dimensional subspaces (Nesterov ’11, Gratton et al ’15, etc).
- Recent: choose those directions in random subspaces (Kozak et al ’21,’22 for directional derivative estimates).
Recall: Classical direct search

- Set $\mathcal{D}_k \subset \mathbb{R}^n$, $|\mathcal{D}_k| = m$, $\text{cm}(\mathcal{D}_k) \geq \kappa$;
- Complexity:
  \[ \mathcal{O}(m\kappa^{-2}\epsilon^{-2}) \].

- $m$ may not depend on $n$ (probabilistic)
- ...but $\kappa$ depends on $n$ (approximate $\nabla f(x_k) \in \mathbb{R}^n$).
Recall: Classical direct search

- Set $\mathcal{D}_k \subset \mathbb{R}^n$, $|\mathcal{D}_k| = m$, $\text{cm}(\mathcal{D}_k) \geq \kappa$;
- Complexity:

  \[ \mathcal{O}(m\kappa^{-2} \epsilon^{-2}) \].

- $m$ may not depend on $n$ (probabilistic)
- ...but $\kappa$ depends on $n$ (approximate $\nabla f(x_k) \in \mathbb{R}^n$).

My original thought

- Generate directions in a random subspace of $\mathbb{R}^n$;
- Use results from dimensionality reduction;
- Remove all dependencies on $n$!
Recall: Classical direct search

- Set $\mathcal{D}_k \subset \mathbb{R}^n$, $|\mathcal{D}_k| = m$, $\text{cm}(\mathcal{D}_k) \geq \kappa$;
- Complexity:
  $$\mathcal{O}(m\kappa^{-2} \epsilon^{-2}).$$

- $m$ may not depend on $n$ (probabilistic)
- ...but $\kappa$ depends on $n$ (approximate $\nabla f(x_k) \in \mathbb{R}^n$).

My original thought

- Generate directions in a random subspace of $\mathbb{R}^n$;
- Use results from dimensionality reduction;
- Remove all dependencies on $n$!

**Spoiler alert: You cannot do that!**
What can you do?

**Approach**
- Consider a random subspace of dimension $r \leq n$;
- Use a PSS to approximate the projected gradient in the subspace;
- Guarantee sufficient gradient information *in probability*.

**What it brings us**
- Handle *unbounded directions*;
- Revisit the opposite uniform directions choice;
- Generalize the analysis to other settings, e.g. Gaussian.
**Inputs:** \( x_0 \in \mathbb{R}^n, \alpha_0 > 0, 0 < \gamma_{\text{dec}} < 1 < \gamma_{\text{inc}}. \)

**Iteration \( k \):** Given \((x_k, \alpha_k)\),

- Choose \( P_k \in \mathbb{R}^{r \times n} \) **at random**.
- Choose \( D_k \subset \mathbb{R}^r \) having \( m \) vectors.
- If \( \exists \, d_k \in D_k \) such that
  \[
  f(x_k + \alpha_k P_k^T d_k) < f(x_k) - \alpha_k^2 \| P_k^T d_k \|^2,
  \]
  set \( x_{k+1} := x_k + \alpha_k P_k^T d_k, \) \( \alpha_{k+1} := \gamma_{\text{inc}} \alpha_k \).
- Otherwise, set \( x_{k+1} := x_k, \) \( \alpha_{k+1} := \gamma_{\text{dec}} \alpha_k \).
Probabilistic properties

New polling sets

\[ \{ P_k^T d \mid d \in \mathcal{D}_k \} \subset \mathbb{R}^n. \]

- \( P_k \in \mathbb{R}^{r \times n} \): Maps onto \( r \)-dimensional subspace;
- \( \mathcal{D}_k \): Direction set in \( \mathbb{R}^r \).

What do we want?

- Preserve information while applying \( P_k / P_k^T \).
- Approximate \(-P_k \nabla f(x_k)\) using \( \mathcal{D}_k \).
$P_k$ is $(\eta, \sigma, P_{\text{max}})$-well aligned for $(f, x_k)$ if

$$
\begin{align*}
\| P_k \nabla f(x_k) \| & \geq \eta \| \nabla f(x_k) \|, \\
\sigma_{\text{min}}(P_k) & \geq \sigma, \\
\sigma_{\text{max}}(P_k) & \leq P_{\text{max}}.
\end{align*}
$$
Probabilistic properties for $P_k$

$P_k$ is $(\eta, \sigma, P_{\text{max}})$-well aligned for $(f, x_k)$ if

$$
\begin{align*}
\|P_k \nabla f(x_k)\| & \geq \eta \|\nabla f(x_k)\|, \\
\sigma_{\min}(P_k) & \geq \sigma, \\
\sigma_{\max}(P_k) & \leq P_{\text{max}}.
\end{align*}
$$

Ex) $P_k = I_n \in \mathbb{R}^{n \times n}$ is $(1, 1, 1)$-well aligned.
Probabilistic properties for $P_k$

$P_k$ is $(\eta, \sigma, P_{\text{max}})$-well aligned for $(f, x_k)$ if

$$
\begin{cases}
\|P_k \nabla f(x_k)\| \geq \eta \|\nabla f(x_k)\|, \\
\sigma_{\text{min}}(P_k) \geq \sigma, \\
\sigma_{\text{max}}(P_k) \leq P_{\text{max}}.
\end{cases}
$$

Ex) $P_k = I_n \in \mathbb{R}^{n \times n}$ is $(1, 1, 1)$-well aligned.

Probabilistic version

$\{P_k\}$ is $(q, \eta, \sigma, P_{\text{max}})$-well aligned if:

$$
\mathbb{P}(P_0 \text{ $(q, \eta, \sigma, P_{\text{max}})$-well aligned }) \geq q
$$

$$
\forall k \geq 1, \quad \mathbb{P}((q, \eta, \sigma, P_{\text{max}})\text{-well aligned } \mid P_0, D_0, \ldots, P_{k-1}, D_{k-1}) \geq q,
$$
Deterministic descent

The set $\mathcal{D}_k$ is $(\kappa, d_{\text{max}})$-descent for $(f, x_k)$ if

$$\max_{d \in \mathcal{D}_k} \frac{-d^T P_k \nabla f(x_k)}{\|d\| \|P_k \nabla f(x_k)\|} \geq \kappa,$$

$$\forall d \in \mathcal{D}_k, \quad d_{\text{max}}^{-1} \leq \|d\| \leq d_{\text{max}}.$$
Deterministic descent

The set $\mathcal{D}_k$ is $(\kappa, d_{\text{max}})$-descent for $(f, x_k)$ if

$$\max_{d \in \mathcal{D}_k} \frac{-d^T P_k \nabla f(x_k)}{\|d\| \|P_k \nabla f(x_k)\|} \geq \kappa,$$

$$\forall d \in \mathcal{D}_k, \quad d_{-1} \leq \|d\| \leq d_{\text{max}}.$$

Ex) $D_\oplus = \{e_1, \ldots, e_n, -e_1, \ldots, -e_n\}$ is $(\frac{1}{\sqrt{n}}, 1)$-descent.
Deterministic descent

The set \( \mathcal{D}_k \) is \((\kappa, d_{\text{max}})\)-descent for \((f, x_k)\) if

\[
\begin{aligned}
\max_{d \in \mathcal{D}_k} \frac{-d^T P_k \nabla f(x_k)}{\|d\| \|P_k \nabla f(x_k)\|} & \geq \kappa, \\
\forall d \in \mathcal{D}_k, \quad d_{\text{max}}^{-1} \leq \|d\| \leq d_{\text{max}}.
\end{aligned}
\]

Ex) \( D_\oplus = \{e_1, \ldots, e_n, -e_1, \ldots, -e_n\} \) is \((\frac{1}{\sqrt{n}}, 1)\)-descent.

Probabilistic descent sets

\{\mathcal{D}_k\} is \((p, \kappa, d_{\text{max}})\)-descent if:

\[
\mathbb{P} (\mathcal{D}_0 (\kappa, d_{\text{max}})\text{-descent} \mid P_0) \geq p
\]

\[
\forall k \geq 1, \quad \mathbb{P} (\mathcal{D}_k (\kappa, d_{\text{max}})\text{-descent} \mid P_0, \mathcal{D}_0, \ldots, P_{k-1}, \mathcal{D}_{k-1}, P_k) \geq p,
\]
Key arguments

Small step size + Good $P_k/D_k \Rightarrow$ Success

If $P_k$ is $(\eta, \sigma, P_{\text{max}})$-well aligned, $D_k$ is $(\kappa, d_{\text{max}})$-descent, and

$$\alpha_k < O\left(\frac{\kappa \eta}{P_{\text{max}}^2 d_{\text{max}}^3} \|\nabla f(x_k)\|\right).$$

then $x_{k+1} \neq x_k$ and $\alpha_{k+1} \geq \alpha_k$. 

For all realizations of the method, $X_k \in K$ if $\alpha_k^2 k \leq O \text{d}^2 \sigma^2 \text{< } \infty$, where $K$ is the set of successful iterations for which $P_k$ is $(\eta, \sigma, P_{\text{max}})$-well aligned and $D_k$ is $(\kappa, d_{\text{max}})$-descent.
Key arguments

Small step size + Good $P_k/D_k \Rightarrow$ Success

If $P_k$ is $(\eta, \sigma, P_{\text{max}})$-well aligned, $D_k$ is $(\kappa, d_{\text{max}})$-descent, and

$$\alpha_k < \mathcal{O} \left( \frac{\kappa \eta}{P_{\text{max}}^2 d_{\text{max}}^3} \| \nabla f(x_k) \| \right).$$

then $x_{k+1} \neq x_k$ and $\alpha_{k+1} \geq \alpha_k$.

A step size sequence goes to zero

For all realizations of the method,

$$\sum_{k \in \mathcal{K}} \alpha_k^2 \leq \mathcal{O} \left( \frac{d_{\text{max}}^2}{\sigma^2} \right) < \infty,$$

where $\mathcal{K}$ is the set of successful iterations for which $P_k$ is $(\eta, \sigma, P_{\text{max}})$-well aligned and $D_k$ is $(\kappa, d_{\text{max}})$-descent.
Theorem (Roberts and Royer, 2022)

Assume:

- $\{D_k\}$ $(p, \kappa, d_{\text{max}})$-descent, $|D_k| = m$;
- $\{P_k\}$ $(q, \eta, \sigma, P_{\text{max}})$-well aligned.

Let $N_\epsilon$ the number of function evaluations needed to have $\|\nabla f(x_k)\| \leq \epsilon$.

$$
\mathbb{P} \left( N_\epsilon \leq \mathcal{O} \left( \frac{m\phi \epsilon^{-2}}{pq - p_0} \right) \right) \geq 1 - \exp \left( -\mathcal{O} \left( \frac{pq - p_0}{pq} \phi \epsilon^{-2} \right) \right).
$$

where $\phi = \eta^{-2} \sigma^{-2} P_{\text{max}}^4 d_{\text{max}}^8 \kappa^{-2}$.
Theorem (Roberts and Royer, 2022)

Assume:

- \( \{D_k\} \) \((p, \kappa, d_{\text{max}})\)-descent, \(|D_k| = m\);
- \( \{P_k\} \) \((q, \eta, \sigma, P_{\text{max}})\)-well aligned.

Let \( N_\epsilon \) the number of function evaluations needed to have \( \|\nabla f(x_k)\| \leq \epsilon \).

\[
\mathbb{P} \left( N_\epsilon \leq O \left( \frac{m\phi \epsilon^{-2}}{pq - p_0} \right) \right) \geq 1 - \exp \left( -O \left( \frac{pq - p_0}{pq} \phi \epsilon^{-2} \right) \right).
\]

where \( \phi = \eta^{-2}\sigma^{-2} P_{\text{max}}^4 d_{\text{max}}^8 \kappa^{-2} \).

Does this bound depend on \( n \)?
Can we really improve the dimension dependence?

\[ m \eta^{-2} \sigma^{-2} P^4_{\text{max}} d^8_{\text{max}} \kappa^{-2} \epsilon^{-2}. \]
Can we really improve the dimension dependence?

\[ m \eta^{-2} \sigma^{-2} P_{\text{max}}^4 d_{\text{max}}^8 \kappa^{-2} \epsilon^{-2}. \]

A first simplification

- \( D_k = \{ e_1, \ldots, e_r, -e_1, \ldots, -e_r \} \) in \( \mathbb{R}^r \);
- \( \kappa = \frac{1}{\sqrt{r}} \), \( m = 2r \), \( d_{\text{max}} = 1 \).

\( \Rightarrow \) Bound becomes \( 2r^2 \eta^{-2} \sigma^{-2} P_{\text{max}}^4 \epsilon^{-2} \).
Can we really improve the dimension dependence?

\[ m \eta^{-2} \sigma^{-2} P_{\text{max}}^4 d_{\text{max}}^8 \kappa^{-2} \epsilon^{-2}. \]

A first simplification

- \( D_k = \{e_1, \ldots, e_r, -e_1, \ldots, -e_r\} \) in \( \mathbb{R}^r \);
- \( \kappa = \frac{1}{\sqrt{r}}, \; m = 2r, \; d_{\text{max}} = 1. \)

⇒ Bound becomes \( 2r^2 \eta^{-2} \sigma^{-2} P_{\text{max}}^4 \epsilon^{-2} \).

Using sketching techniques

<table>
<thead>
<tr>
<th>( P_k )</th>
<th>( \sigma )</th>
<th>( P_{\text{max}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identity</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Gaussian</td>
<td>( \Theta(\sqrt{n/r}) )</td>
<td>( \Theta(\sqrt{n/r}) )</td>
</tr>
<tr>
<td>Hashing</td>
<td>( \Theta(\sqrt{n/r}) ) (Dzahini &amp; Wild ’22b)</td>
<td>( \sqrt{n} )</td>
</tr>
<tr>
<td>Orthogonal</td>
<td>( \sqrt{n/r} )</td>
<td>( \sqrt{n/r} ).</td>
</tr>
</tbody>
</table>

⇒ Get a bound in \( O(n \epsilon^{-2}) \) even when \( r = O(1) \) and \( \eta = O(1)! \)
1 Direct search

2 Probabilistic descent

3 In reduced subspaces

4 It works
Experiments in large dimensions

**Benchmark:**
- Medium-scale test set (90 CUTEst problems of dimension $\approx 100$);
- Large-scale test set (28 CUTEst problems of dimension $\approx 1000$).

**Budget:** $200(n + 1)$ evaluations.

**Comparison:**
- Deterministic DS with $D_k = D_\oplus$ or $D_k = \{e_1, \ldots, e_n, -\sum_{i=1}^{n} e_i\}$;
- Probabilistic direct search with 2 uniform directions;
- Stochastic Three Point;
- Probabilistic direct search with Gaussian/Hashing/Orthogonal $P_k$ matrices + 2 directions in the subspace.

**Goal:** Satisfy $f(x_k) - f_{opt} \leq 0.1(f(x_0) - f_{opt})$. 
Comparison of all methods

Left: Medium scale; Right: Large scale.

- Can use less directions through sketching;
- But always a (hidden) dependency on \( n \)!
Gaussian matrices and the value of $r$

Left: Medium scale; Right: Large scale.

Numerically

- Sketches of dimension $> 1$ may improve things...
- ...but in general opposite (Gaussian) directions work best!
More on numerics (feat. A. L. Custódio, E. Silva)

An email exchange last April

- Issue with replicating the results;
- Turns out the settings were not identical.
An email exchange last April

- Issue with replicating the results;
- Turns out the settings were not identical.

Implementation details

- **Increasing the step size on successful iterations is key to performance**
  - Required for our theory;
  - Not used in stochastic three point.

- **Sufficient decrease less critical**

  *In our experiments, we used*

  \[
  f(x_k + \alpha_k d_k) < f(x_k) - \min\{10^{-3}, 10^{-3} \alpha_k^2 \|d_k\|^2\}.
  \]
Our findings

- A revised probabilistic analysis/subspace viewpoint;
- Allows for unbounded directions;
- Complexity (still) in $\mathcal{O}(n)$. 

"But why do 1D subspaces work?" (W. Hare)
Summary

Our findings
- A revised probabilistic analysis/subspace viewpoint;
- Allows for unbounded directions;
- Complexity (still) in $O(n)$.

Perspectives
- Stochastic function values.
- “But why do 1D subspaces work?” (W. Hare)
That’s it!

The paper

Direct search based on probabilistic descent in reduced spaces

The package

- https://github.com/lindonroberts/directsearch
- In Python, has all experiments.

C. W. Royer
Direct search in reduced subspaces
That’s it!

The paper

Direct search based on probabilistic descent in reduced spaces

The package

- https://github.com/lindonroberts/directsearch
- In Python, has all experiments.

Thank you for your attention!
clement.royer@lamsade.dauphine.fr