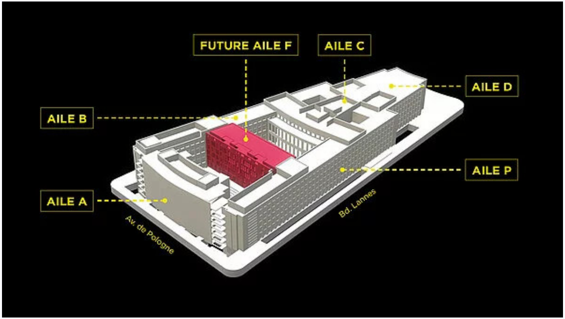
A derivative-free algorithm resilient to straggler function evaluations

Clément W. Royer

Based on joint works with W. Hare, G. Jarry-Bolduc, S. Kerleau

Derivative-Free Optimization Symposium - Padova, June 24, 2024

Motivation: Dauphine's *Nouveau Campus*



- New wing in construction ⇒ 2024.
- Others renovated in order: B, P, C+D, A.
- Expected year of completion: 2027.

Our task: Allocate office space during the renovation process.

Motivation: Dauphine's *Nouveau Campus* ('ed)

Our model for the Dauphine problem

- Huge integer LP, solved via Gurobi.
- ~ 20 hyperparameters defining the model.
- Parallel runs on the department server.

Sub-task: Optimize hyperparameters.

Motivation: Dauphine's *Nouveau Campus* ('ed)

Our model for the Dauphine problem

- Huge integer LP, solved via Gurobi.
- $\sim$ 20 hyperparameters defining the model.
- Parallel runs on the department server.

Sub-task: Optimize hyperparameters.

Problem challenges

- Cannot differentiate (easily) within Gurobi
  $\Rightarrow$ Derivative-free algorithms!
- Solving time depends on hyperparameters (3-48 hours for a feasible point!)
  $\Rightarrow$ Limited benefits of parallelism.

Motivation: Dauphine's *Nouveau Campus* ('ed)

Our model for the Dauphine problem

- Huge integer LP, solved via Gurobi.
- $\sim$ 20 hyperparameters defining the model.
- Parallel runs on the department server.

Sub-task: Optimize hyperparameters.

Problem challenges

- Cannot differentiate (easily) within Gurobi
  $\Rightarrow$ Derivative-free algorithms!
- Solving time depends on hyperparameters (3-48 hours for a feasible point!)
  $\Rightarrow$ Limited benefits of parallelism.
- Department chair wants weekly updates/new tests
  $\Rightarrow$ Need to give up on some tests.

Motivation: Dauphine's *Nouveau Campus* ('ed)

Our model for the Dauphine problem

- Huge integer LP, solved via Gurobi.
- $\sim$ 20 hyperparameters defining the model.
- Parallel runs on the department server.

Sub-task: Optimize hyperparameters.

Problem challenges

- Cannot differentiate (easily) within Gurobi
  $\Rightarrow$ Derivative-free algorithms!
- Solving time depends on hyperparameters (3-48 hours for a feasible point!)
  $\Rightarrow$ Limited benefits of parallelism.
- Department chair wants weekly updates/new tests
  $\Rightarrow$ Need to give up on some tests.

Motivation: Dauphine's *Nouveau Campus* ('ed)

Our model for the Dauphine problem

- Huge integer LP, solved via Gurobi.
- $\sim$ 20 hyperparameters defining the model.
- Parallel runs on the department server.

Sub-task: Optimize hyperparameters.

Problem challenges

- Cannot differentiate (easily) within Gurobi
  $\Rightarrow$ Derivative-free algorithms!
- Solving time depends on hyperparameters (3-48 hours for a feasible point!)
  $\Rightarrow$ Limited benefits of parallelism.
- Department chair wants weekly updates/new tests
  $\Rightarrow$ Need to give up on some tests.

Can we cook up an algorithm adapted to this setting?

Outline

1. Formal problem and algorithm
2. Building and using PkSS
3. Conclusion

Outline

1. Formal problem and algorithm
2. Building and using PkSS
3. Conclusion

Optimization problem

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

- f bounded below, nonconvex, $\mathcal{C}^{1,1}$ (for analysis).

Optimization problem

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

- f bounded below, nonconvex, $\mathcal{C}^{1,1}$ (for analysis).
- Derivatives unavailable for algorithmic use.

Optimization problem

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

- f bounded below, nonconvex, $\mathcal{C}^{1,1}$ (for analysis).
- Derivatives unavailable for algorithmic use.
- Parallel evaluations of f allowed.

Optimization problem

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

- f bounded below, nonconvex, $\mathcal{C}^{1,1}$ (for analysis).
- Derivatives unavailable for algorithmic use.
- **Parallel evaluations of f allowed**.
- Evaluations can take unusually long⇒**Stragglers**.

A (simplified) direct-search framework

Inputs: $x_0 \in \mathbb{R}^n$, $\alpha_0 > 0$.

Iteration i**:** Given (x_i, α_i),

- Choose a set $\mathcal{D}_i \subset \mathbb{R}^n$ of m (nonzero) vectors.
- If $\exists\ d_i \in \mathcal{D}_i$ such that

$$f(x_i + \alpha_i\, d_i) < f(x_i) - \alpha_i^2 \|d_i\|^2$$

  set $x_{i+1} := x_i + \alpha_i d_i$, $\alpha_{i+1} := 2\alpha_i$.
- Otherwise, set $x_{i+1} := x_i$, $\alpha_{i+1} := \alpha_i/2$.

A (simplified) direct-search framework

Inputs: $x_0 \in \mathbb{R}^n$, $\alpha_0 > 0$.
Iteration i**:** Given (x_i, α_i),

- Choose a set $\mathcal{D}_i \subset \mathbb{R}^n$ of m (nonzero) vectors.
- If $\exists\ d_i \in \mathcal{D}_i$ such that

$$f(x_i + \alpha_i\, d_i) < f(x_i) - \alpha_i^2 \|d_i\|^2$$

  set $x_{i+1} := x_i + \alpha_i d_i$, $\alpha_{i+1} := 2\alpha_i$.
- Otherwise, set $x_{i+1} := x_i$, $\alpha_{i+1} := \alpha_i/2$.

- **Key**: Choice of $\mathcal{D}_i$.
- **Parallel version**: Works as long as all directions in $\mathcal{D}_i$ have been polled when $x_{i+1} = x_i$.

Classical direct-search theory

Usual tool: Positive spanning sets (PSS)

- $\mathcal{D} \subset \mathbb{R}^n$ PSS if it spans $\mathbb{R}^n$ by nonnegative linear combinations
  $\Rightarrow \; |\mathcal{D}| \geq n + 1$.

Classical direct-search theory

Usual tool: Positive spanning sets (PSS)

- $\mathcal{D} \subset \mathbb{R}^n$ PSS if it spans $\mathbb{R}^n$ by nonnegative linear combinations
  $\Rightarrow \ |\mathcal{D}| \geq n+1$.
- $\mathcal{D}$ positive basis if no proper subset of $\mathcal{D}$ is a PSS
  $\Rightarrow \ n+1 \leq |\mathcal{D}| \leq 2n$.

Classical direct-search theory

Usual tool: Positive spanning sets (PSS)

- $\mathcal{D} \subset \mathbb{R}^n$ PSS if it spans $\mathbb{R}^n$ by **nonnegative linear combinations**
  $\Rightarrow \ |\mathcal{D}| \geq n+1$.
- $\mathcal{D}$ **positive basis** if no proper subset of $\mathcal{D}$ is a PSS
  $\Rightarrow \ n+1 \leq |\mathcal{D}| \leq 2n$.

Cosine measure

- For any $\mathcal{D} \subset \mathbb{R}^n$, the cosine measure of $\mathcal{D}$ is

$$\mathrm{cm}(\mathcal{D}) := \min_{v \neq 0} \max_{d \in \mathcal{D}} \frac{v^{\mathrm{T}} d}{\|v\| \|d\|}.$$

- $\mathcal{D}$ PSS $\Longleftrightarrow \mathrm{cm}(\mathcal{D}) > 0$.

Classical direct-search theory ('ed)

Theorem: Complexity of direct search

Apply direct search with $\mathcal{D}_i = \mathcal{D}\ \forall i$, $\mathcal{D}$ PSS. Then the method satisfies

$$\min_{0 \leq i \leq J} \|\nabla f(x_i)\| \leq \epsilon$$

in at most

$$J = \mathcal{O}\left(|\mathcal{D}|\, \mathrm{cm}(\mathcal{D})^{-2} \epsilon^{-2}\right)$$

function evaluations.

Classical direct-search theory ('ed)

Theorem: Complexity of direct search

Apply direct search with $\mathcal{D}_i = \mathcal{D} \; \forall i$, $\mathcal{D}$ PSS. Then the method satisfies

$$\min_{0 \le i \le J} \|\nabla f(x_i)\| \le \epsilon$$

in at most

$$J = \mathcal{O}\left(|\mathcal{D}| \, \mathrm{cm}(\mathcal{D})^{-2} \epsilon^{-2}\right)$$

function evaluations.

Typical values

- $|\mathcal{D}| = \mathcal{O}(n)$.
- $\mathrm{cm}(\mathcal{D}) = \mathcal{O}(n^a)$, $a \in \{-0.5, -1\}$.

Direct search and stragglers

Inputs: $x_0 \in \mathbb{R}^n$, $\alpha_0 > 0$.
Iteration i: Given (x_i, α_i),

- Choose a set $\mathcal{D}_i \subset \mathbb{R}^n$ of m (nonzero) vectors.
- If $\exists\ d_i \in \mathcal{D}_i$ such that

$$f(x_i + \alpha_i\, d_i) < f(x_i) - \alpha_i^2 \|d_i\|^2$$

  set $x_{i+1} := x_i + \alpha_i d_i$, $\alpha_{i+1} := 2\alpha_i$.
- Otherwise, set $x_{i+1} := x_i$, $\alpha_{i+1} := \alpha_i/2$.

Direct search and stragglers

Inputs: $x_0 \in \mathbb{R}^n$, $\alpha_0 > 0$.

Iteration i**:** Given (x_i, α_i),

- Choose a set $\mathcal{D}_i \subset \mathbb{R}^n$ of m (nonzero) vectors.
- If $\exists\ d_i \in \mathcal{D}_i \setminus \mathcal{S}_i$ such that

$$f(x_i + \alpha_i\, d_i) < f(x_i) - \alpha_i^2 \|d_i\|^2$$

  set $x_{i+1} := x_i + \alpha_i d_i$, $\alpha_{i+1} := 2\alpha_i$.
- Otherwise, set $x_{i+1} := x_i$, $\alpha_{i+1} := \alpha_i/2$.

Straggler model: At every iteration i,

- $\exists \mathcal{S}_i \subset \mathcal{D}_i$ of **straggler directions** (f much longer to evaluate).
- $\mathcal{S}_i$ unknown before evaluations launched!
- Evaluations in $\mathcal{S}_i$ cannot be used in analyzing the method.

A new PSS property

Positive k-spanning sets (PkSS), $k \geq 1$

- $\mathcal{D} \subset \mathbb{R}^n$ PkSS if any $\mathcal{N} \subset \mathcal{D}$ with $|\mathcal{N}| = |\mathcal{D}| - k + 1$ is a PSS.

A new PSS property

Positive k-spanning sets (PkSS), $k \geq 1$

- $\mathcal{D} \subset \mathbb{R}^n$ PkSS if any $\mathcal{N} \subset \mathcal{D}$ with $|\mathcal{N}| = |\mathcal{D}| - k + 1$ is a PSS.
- $\mathcal{D}$ positive k-basis if no proper subset of $\mathcal{D}$ is a PkSS.

A new PSS property

Positive k-spanning sets (PkSS), $k \geq 1$

- $\mathcal{D} \subset \mathbb{R}^n$ PkSS if any $\mathcal{N} \subset \mathcal{D}$ with $|\mathcal{N}| = |\mathcal{D}| - k + 1$ is a PSS.
- $\mathcal{D}$ positive k-basis if no proper subset of $\mathcal{D}$ is a PkSS.

- Definition inherited from D. A. Marcus ('81, '84).
- $k = 1$: PSS/Positive basis.
- $\mathcal{D}$ PkSS $\Rightarrow$ $|\mathcal{D}| \geq \max\{k, n + 2k - 1\}$.

The k-cosine measure

- For any $\mathcal{D} \subset \mathbb{R}^n$, the k-cosine measure of $\mathcal{D}$ is

$$\mathrm{cm}_k(\mathcal{D}) := \min_{v \neq 0} \max_{\substack{\mathcal{N} \subset \mathcal{D} \\ |\mathcal{N}| = k}} \max_{d \in \mathcal{N}} \frac{v^\mathrm{T} d}{\|v\| \|d\|}.$$

- $\mathcal{D}$ PkSS $\Longleftrightarrow$ $\mathrm{cm}_k(\mathcal{D}) > 0$.

A new cosine measure (Hare, Jarry-Bolduc, Kerleau, R. '24)

The k-cosine measure

- For any $\mathcal{D} \subset \mathbb{R}^n$, the k-cosine measure of $\mathcal{D}$ is

$$\mathrm{cm}_k(\mathcal{D}) := \min_{v \neq 0} \max_{\substack{\mathcal{N} \subset \mathcal{D} \\ |\mathcal{N}|=k}} \max_{d \in \mathcal{N}} \frac{v^{\mathrm{T}} d}{\|v\|\|d\|}.$$

- $\mathcal{D}$ PkSS $\iff \mathrm{cm}_k(\mathcal{D}) > 0$.

Properties

- Equivalent definition:

$$\mathrm{cm}_k(\mathcal{D}) = \min_{\substack{\mathcal{N} \subset \mathcal{D} \\ |\mathcal{N}|=|\mathcal{D}|-k+1}} \mathrm{cm}(\mathcal{N}).$$

- For any $\mathcal{D}$ (not necessarily PkSS !),

$$\mathrm{cm}(\mathcal{D}) \geq \cdots \geq \mathrm{cm}_{|\mathcal{D}|-1}(\mathcal{D}) \geq \mathrm{cm}_{|\mathcal{D}|}(\mathcal{D}).$$

Complexity of direct search

Theorem

- Apply direct search with $\mathcal{D}_i = \mathcal{D}\ \forall i$, $\mathcal{D}$ PkSS.
- Suppose less than k stragglers per iteration: $|\mathcal{S}_i| \le k-1\ \forall i$.

Then the method satisfies

$$\min_{0 \le i \le J} \|\nabla f(x_i)\| \le \epsilon$$

in at most

$$J = \mathcal{O}\left(|\mathcal{D}|\, \mathrm{cm}_k(\mathcal{D})^{-2} \epsilon^{-2}\right)$$

function evaluations.

Theorem

- Apply direct search with $\mathcal{D}_i = \mathcal{D}\ \forall i$, $\mathcal{D}$ PkSS.
- Suppose less than k stragglers per iteration: $|\mathcal{S}_i| \le k-1\ \forall i$.

Then the method satisfies

$$\min_{0 \le i \le J} \|\nabla f(x_i)\| \le \epsilon$$

in at most

$$J = \mathcal{O}\left(|\mathcal{D}|\, \mathrm{cm}_k(\mathcal{D})^{-2} \epsilon^{-2}\right)$$

function evaluations.

Typical values?

- $|\mathcal{D}| \ge \mathcal{O}(n+k)$.
- $\mathrm{cm}_k(\mathcal{D}) = \mathcal{O}(?)$.

Outline

Produce PkSS in practice (1/2)

Our baseline: Maximal (coordinate) positive basis

- $\mathcal{D}_\oplus := \{e_1, \dots, e_n, -e_1, \dots, -e_n\}$, with $\{e_l\}_l$ coordinate basis vectors.
- $\beta\mathcal{D}_\oplus$: multiply all vectors by real β.

Produce PkSS in practice (1/2)

Our baseline: Maximal (coordinate) positive basis

- $\mathcal{D}_{\oplus} := \{e_1, \dots, e_n, -e_1, \dots, -e_n\}$, with $\{e_l\}_l$ coordinate basis vectors.
- $\beta \mathcal{D}_{\oplus}$: multiply all vectors by real β.

First proposal: Duplicate vectors

Let $\beta_1, \dots, \beta_k$ be k distinct positive real numbers. The set

$$\mathcal{D}_{\oplus}^{\beta_{1:k}} := \bigcup_{j=1}^{k} \beta_j \mathcal{D}_{\oplus}$$

is a PkSS with $\mathrm{cm}_k(\mathcal{D}_{\oplus}^{\beta_{1:k}}) = \mathrm{cm}(\mathcal{D}_{\oplus}) = \frac{1}{\sqrt{n}}$.

Produce PkSS in practice (1/2)

Our baseline: Maximal (coordinate) positive basis

- $\mathcal{D}_\oplus := \{e_1, \dots, e_n, -e_1, \dots, -e_n\}$, with $\{e_l\}_l$ coordinate basis vectors.
- $\beta\mathcal{D}_\oplus$: multiply all vectors by real β.

First proposal: Duplicate vectors

Let $\beta_1, \dots, \beta_k$ be k distinct positive real numbers. The set

$$\mathcal{D}_\oplus^{\beta_{1:k}} := \bigcup_{j=1}^{k} \beta_j \mathcal{D}_\oplus$$

is a PkSS with $\mathrm{cm}_k(\mathcal{D}_\oplus^{\beta_{1:k}}) = \mathrm{cm}(\mathcal{D}_\oplus) = \frac{1}{\sqrt{n}}$.

- Easy construction, even yields a positive k-basis!
- Lacks diversity, redundancy if coupled with line search.

Produce PkSS in practice (2/2)

Our baseline: Minimal (coordinate) positive basis

- $\mathcal{D}_{n+1} := \{e_1, \dots, e_n, -\sum_{l=1}^{n} e_l\}$, with $\{e_l\}$ coordinate basis vectors.
- $R\mathcal{D}_{n+1}$: Apply **Rotation matrix** $R \in \mathbb{R}^{n \times n}$ to all vectors.

Produce PkSS in practice (2/2)

Our baseline: Minimal (coordinate) positive basis

- $\mathcal{D}_{n+1} := \{e_1, \dots, e_n, -\sum_{l=1}^{n} e_l\}$, with $\{e_l\}$ coordinate basis vectors.
- $R\mathcal{D}_{n+1}$: Apply **Rotation matrix** $R \in \mathbb{R}^{n \times n}$ to all vectors.

Second proposal: Rotate vectors

Let $R_1, \dots, R_k$ be k distinct positive real numbers. The set

$$\mathcal{D}_{n+1}^{R_{1:k}} := \bigcup_{j=1}^{k} R_j \mathcal{D}_{n+1}$$

is a PkSS with $\mathrm{cm}_k(\mathcal{D}_{n+1}^{R_{1:k}}) \geq \mathrm{cm}(\mathcal{D}_{n+1}) = \frac{1}{\sqrt{n^2+2(n-1)\sqrt{n}}}$.

Produce PkSS in practice (2/2)

Our baseline: Minimal (coordinate) positive basis

- $\mathcal{D}_{n+1} := \{e_1, \dots, e_n, -\sum_{l=1}^{n} e_l\}$, with $\{e_l\}$ coordinate basis vectors.
- $R\mathcal{D}_{n+1}$: Apply **Rotation matrix** $R \in \mathbb{R}^{n \times n}$ to all vectors.

Second proposal: Rotate vectors

Let $R_1, \dots, R_k$ be k distinct positive real numbers. The set

$$\mathcal{D}_{n+1}^{R_{1:k}} := \bigcup_{j=1}^{k} R_j \mathcal{D}_{n+1}$$

is a PkSS with $\mathrm{cm}_k(\mathcal{D}_{n+1}^{R_{1:k}}) \geq \mathrm{cm}(\mathcal{D}_{n+1}) = \frac{1}{\sqrt{n^2+2(n-1)\sqrt{n}}}$.

Includes:

- Positive k-bases/PkSS with duplicates.
- PkSS with no duplicates but not positive k-bases.
- Positive k-basis with no duplicates!

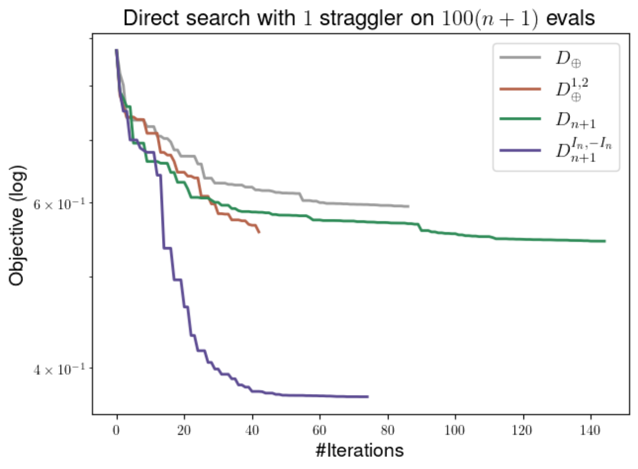
Robust linear regression (Carmon et al '17)

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \frac{1}{2n}\sum_{i=1}^{2n}\phi(a_i^{\mathrm{T}}x-b_i), \quad \phi(t)=\frac{t^2}{1+t^2}.$$

- a_i i.i.d Gaussian, $b_i = a_i^{\mathrm{T}}z + 3u_1 + u_2$, $\{z, u_1\}$ Gaussian, u_2 Bernoulli.

Comparison: Direct search with PSS/P2SS, one straggler/iteration.

$\mathcal{D}$	$\lvert\mathcal{D}\rvert$	$\text{cm}_?$
$\mathcal{D}_\oplus$	$2n$	$\text{cm} = \frac{1}{\sqrt{n}}$
$\mathcal{D}_\oplus^{1,2}$	$4n$	$\text{cm}_2 = \frac{1}{\sqrt{n}}$
$\mathcal{D}_{n+1}$	$n+1$	$\text{cm} = \frac{1}{\sqrt{n^2+2(n-1)\sqrt{n}}}$
$\mathcal{D}_{n+1}^{I_n,-I_n}$	$2n+2$	$\text{cm}_2 = \frac{1}{\sqrt{n^2+2(n-1)\sqrt{n}}}$.



- P2SSs can outperform PSSs with stragglers!
- On 100 runs, $\mathcal{D}_{n+1}^{I_n,-I_n}$ gives best results.

Summing up

DFO with stragglers

- Resilient notion of PSS.
- Convergent algorithm.
- Numerical proof of concept.

References

- W. Hare, G. Jarry-Bolduc, S. Kerleau and C. W. Royer.
  Using orthogonally structured positive bases for constructing positive k-spanning sets with cosine measure guarantees,
  Lin. Alg. Appl. 680:183-207, 2024.
- S. Kerleau and C. W. Royer.
  A derivative-free algorithm resilient to straggler function evaluations,
  Working paper.

Opening for discussions

- About PkSS:
  - Upper bounds on the size of positive k-bases? (Nontrivial!)
  - Best PkSS in terms of ℓ-cosine measure?
  - Connections to strongly connected graphs and neighborly polytopes (!)

Opening for discussions

- About PkSS:
  - Upper bounds on the size of positive k-bases? (Nontrivial!)
  - Best PkSS in terms of ℓ-cosine measure?
  - Connections to strongly connected graphs and neighborly polytopes (!)
- Beyond PkSS:
  - Poisedness equivalent to the PkSS property?
  - Other fun use cases?

Opening for discussions

- About PkSS:
  - Upper bounds on the size of positive k-bases? (Nontrivial!)
  - Best PkSS in terms of ℓ-cosine measure?
  - Connections to strongly connected graphs and neighborly polytopes (!)
- Beyond PkSS:
  - Poisedness equivalent to the PkSS property?
  - Other fun use cases?

Grazie mille!

`clement.royer@lamsade.dauphine.fr`