Direct search based on probabilistic feasible descent for bound and linearly constrained problems

Luis Nunes Vicente
University of Coimbra

FoCM Barcelona, Workshop on Continuous Optimization
July 17, 2017

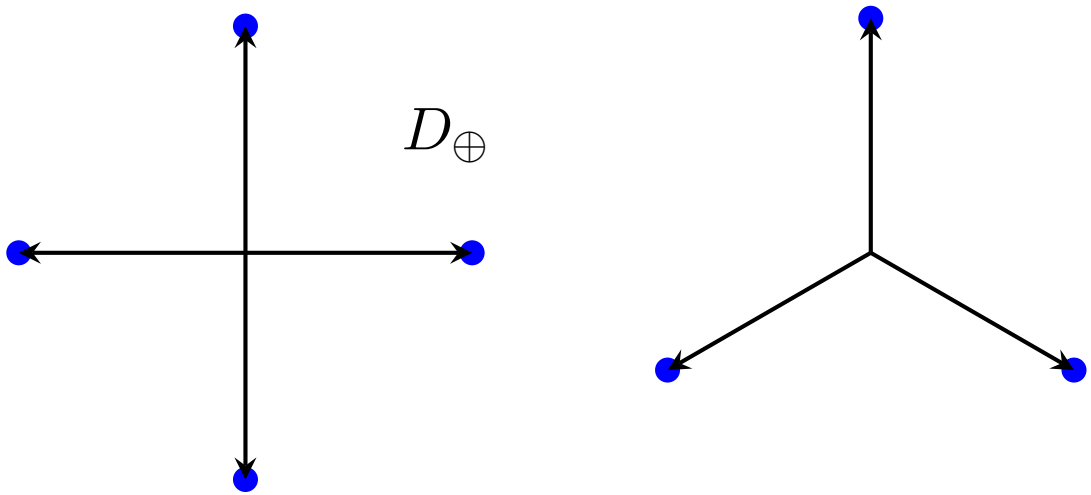
Direct-search methods

Definition

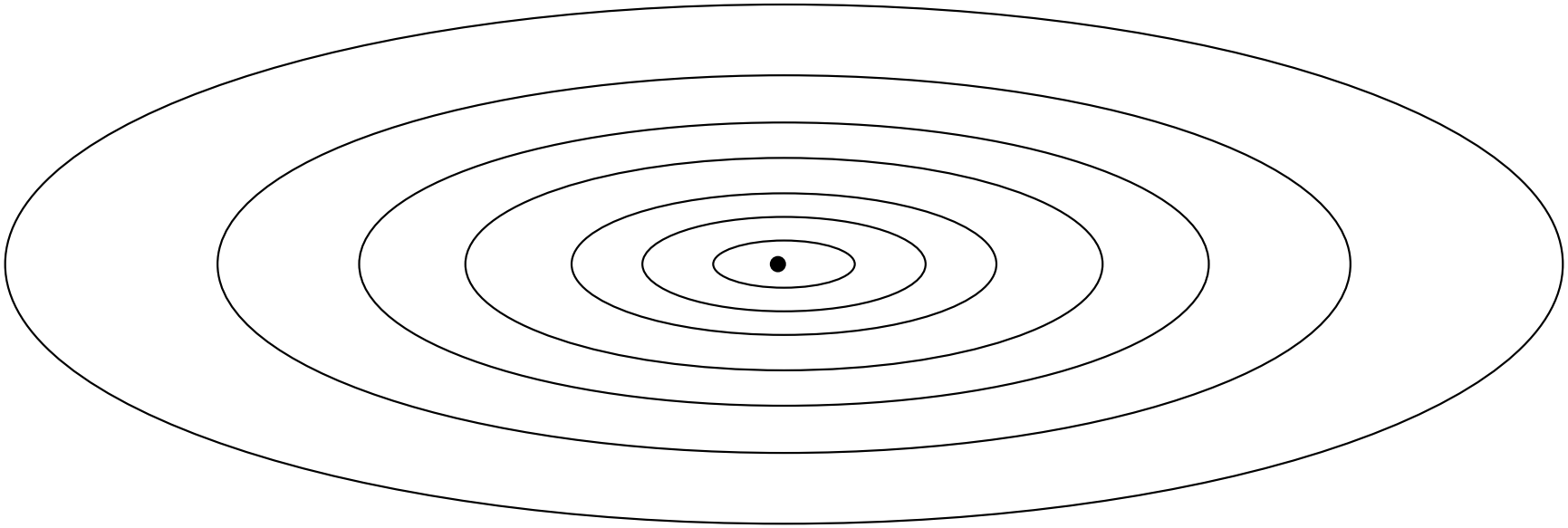
- *Sample the objective function at a finite number of points at each iteration.*
- *Achieve descent by moving in the direction of potentially better points.*
- *In the smooth and deterministic case, these points are defined by directions in positive spanning sets (PSS):*

Definition

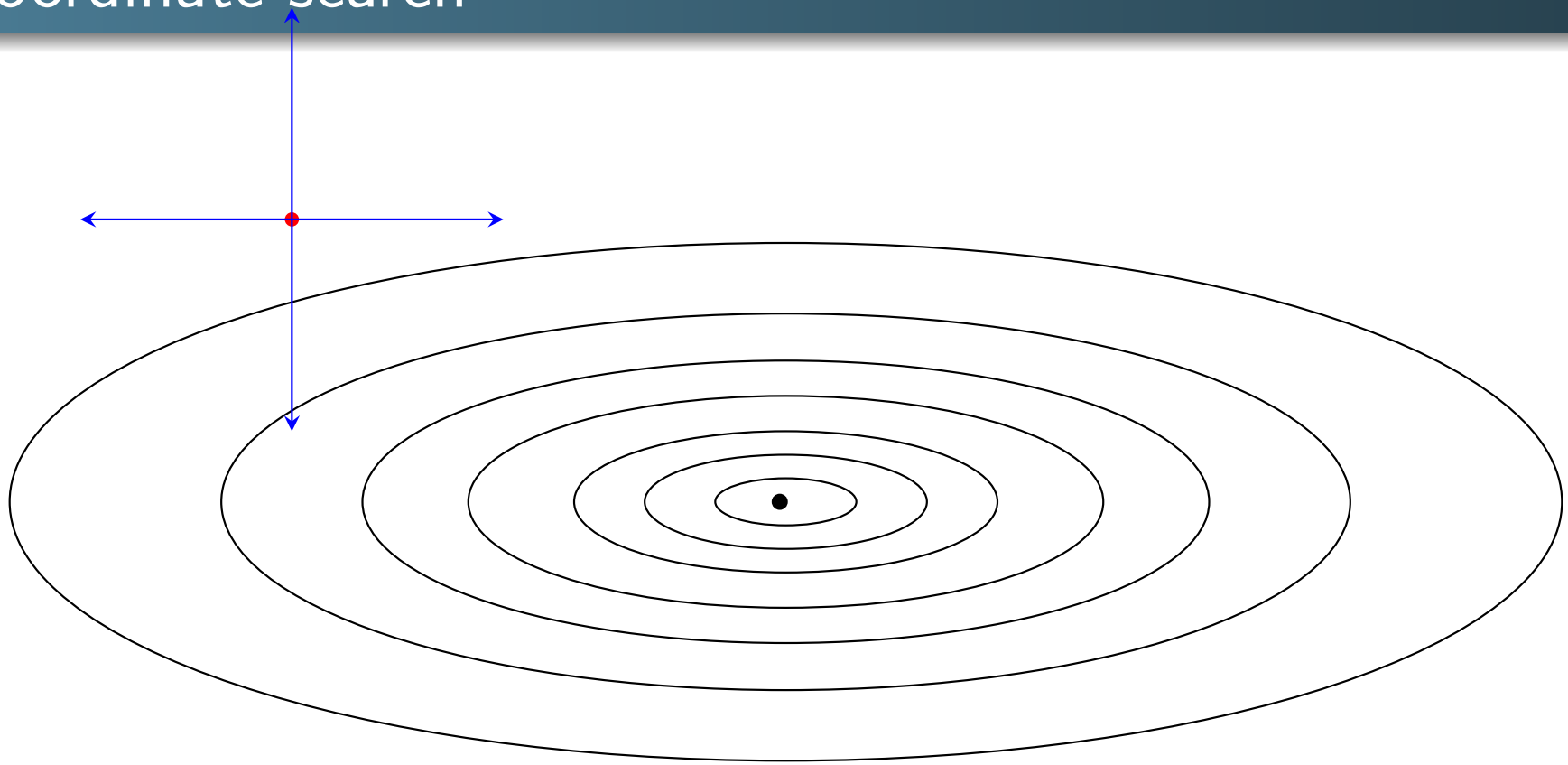
- *Sample the objective function at a finite number of points at each iteration.*
- *Achieve descent by moving in the direction of potentially better points.*
- *In the smooth and deterministic case, these points are defined by directions in positive spanning sets (PSS):*



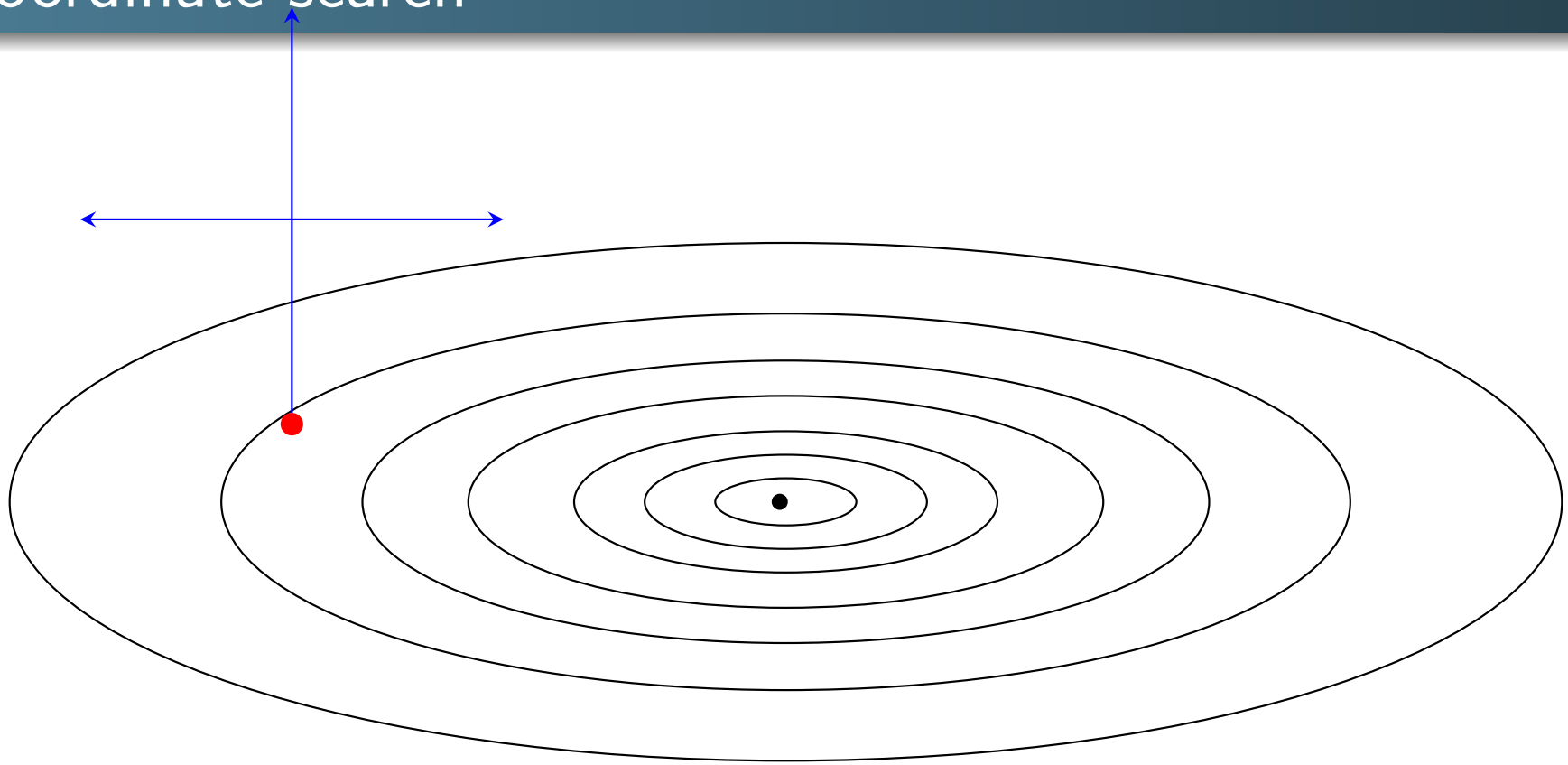
Coordinate search



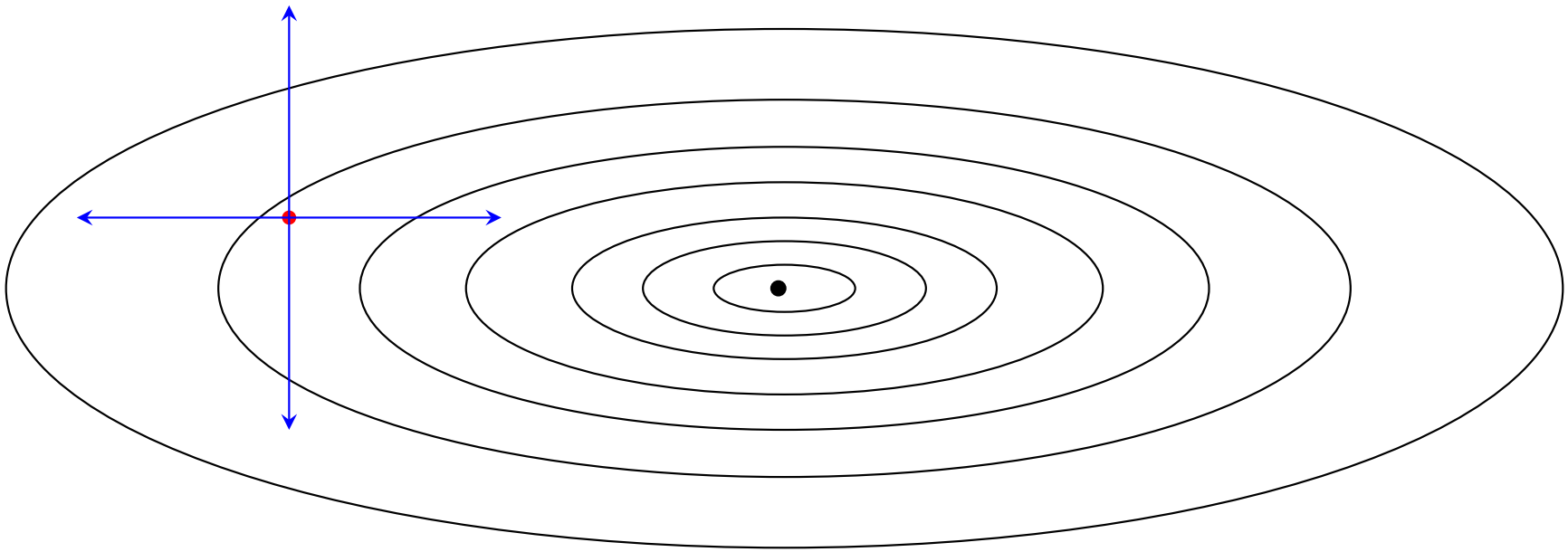
Coordinate search



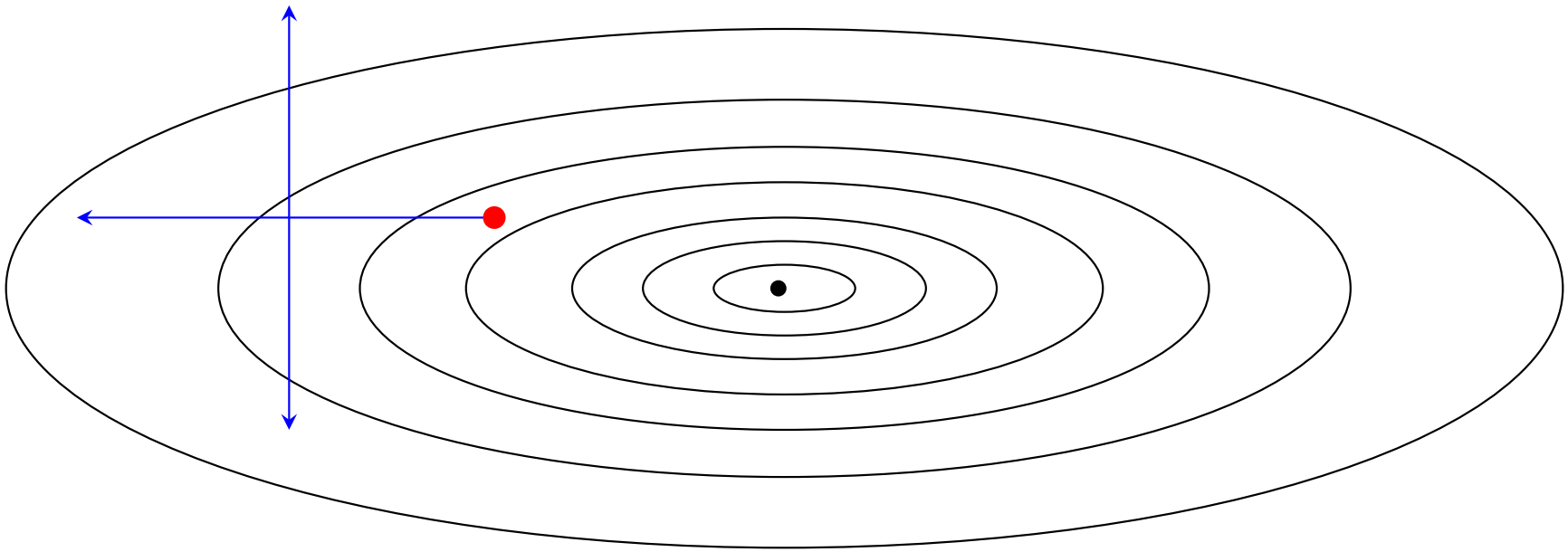
Coordinate search



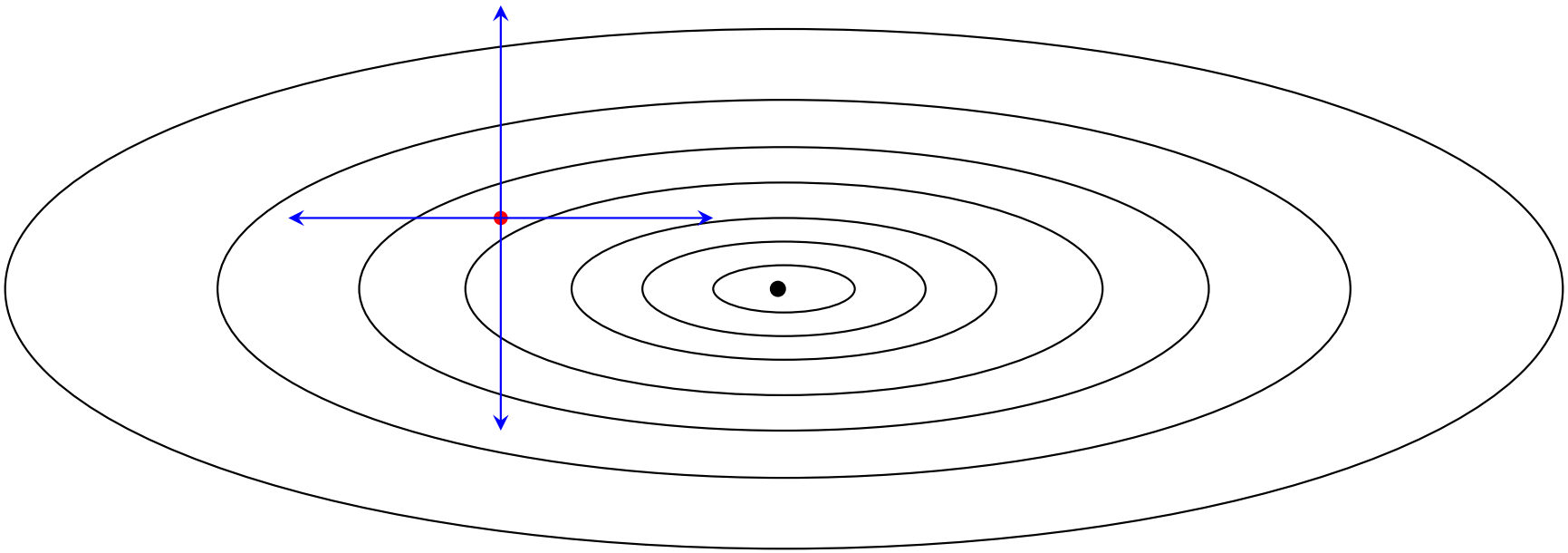
Coordinate search



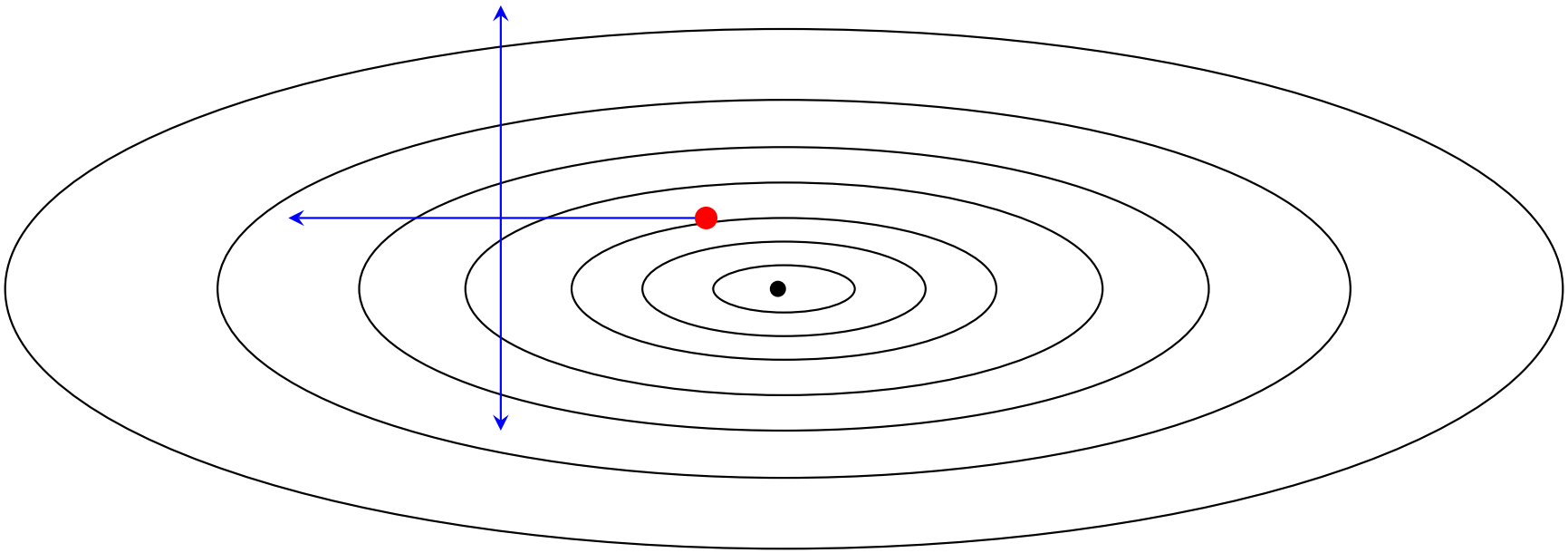
Coordinate search



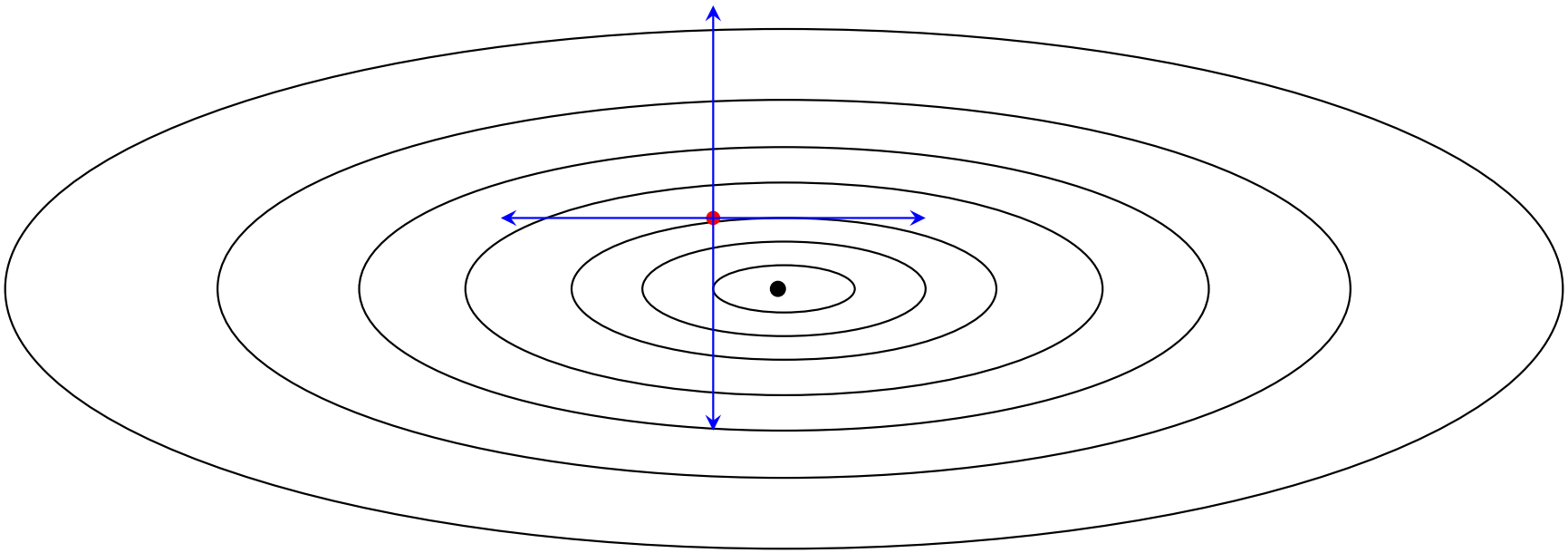
Coordinate search



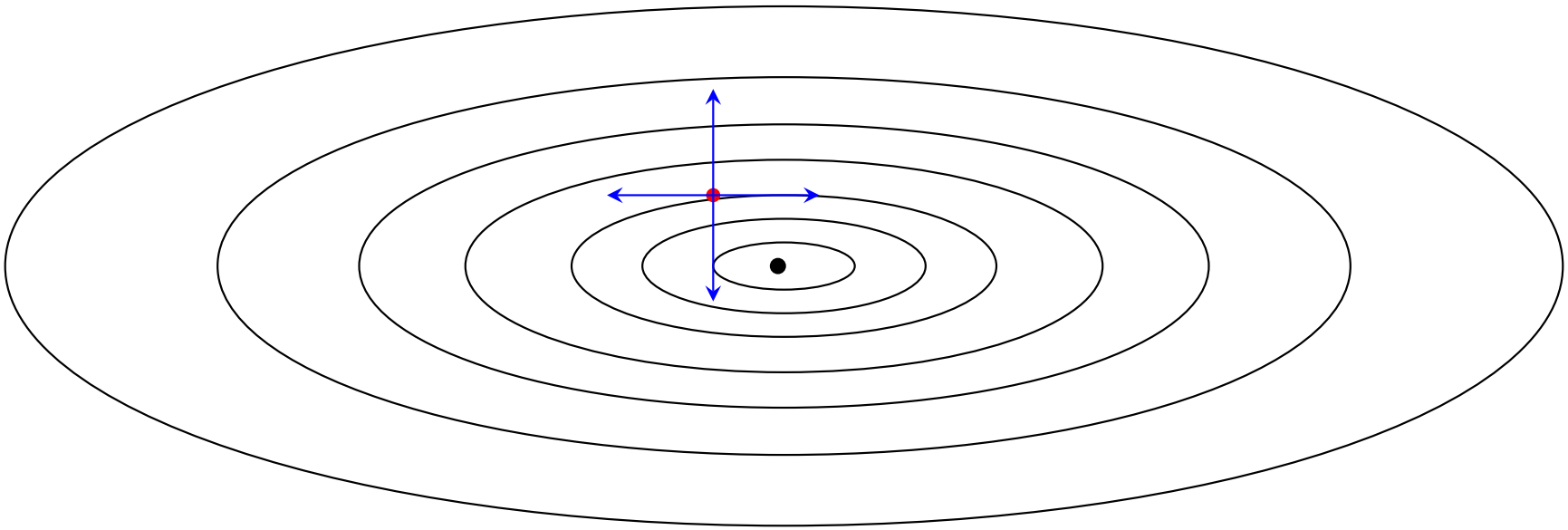
Coordinate search



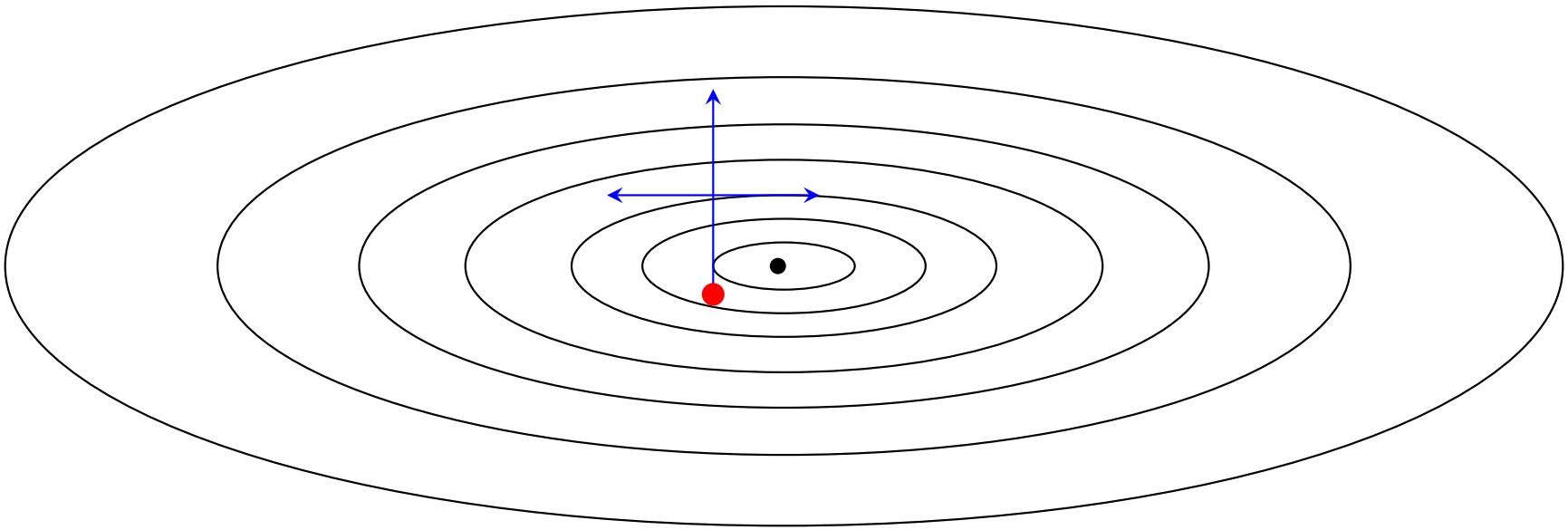
Coordinate search



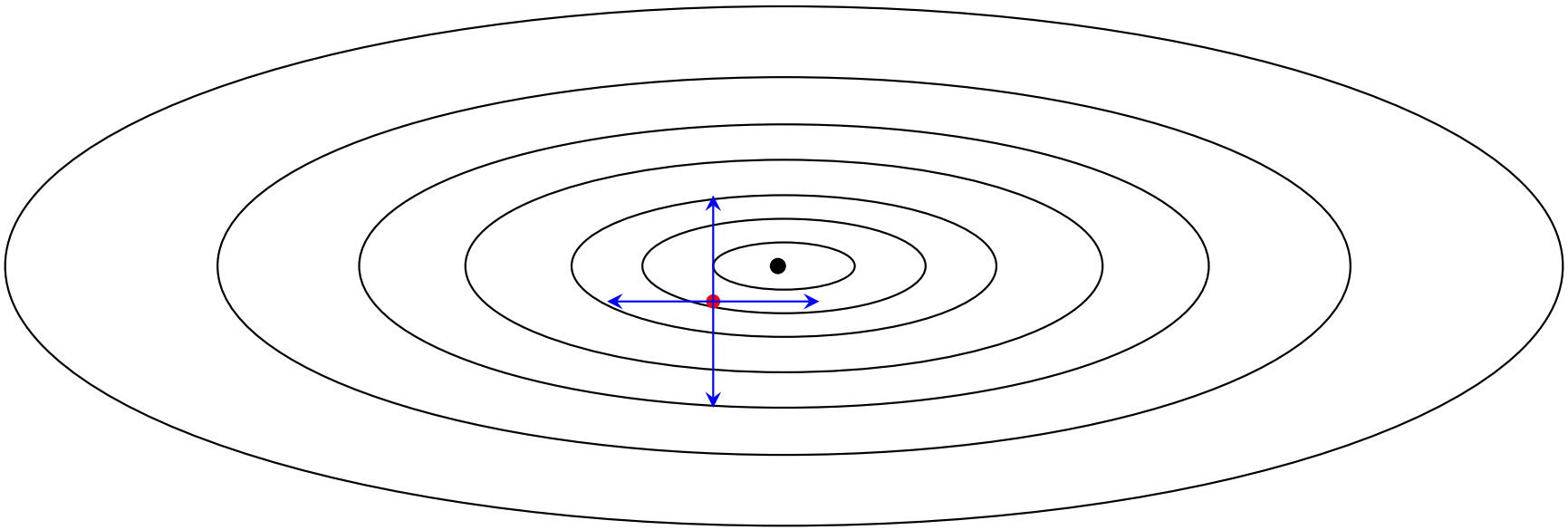
Coordinate search



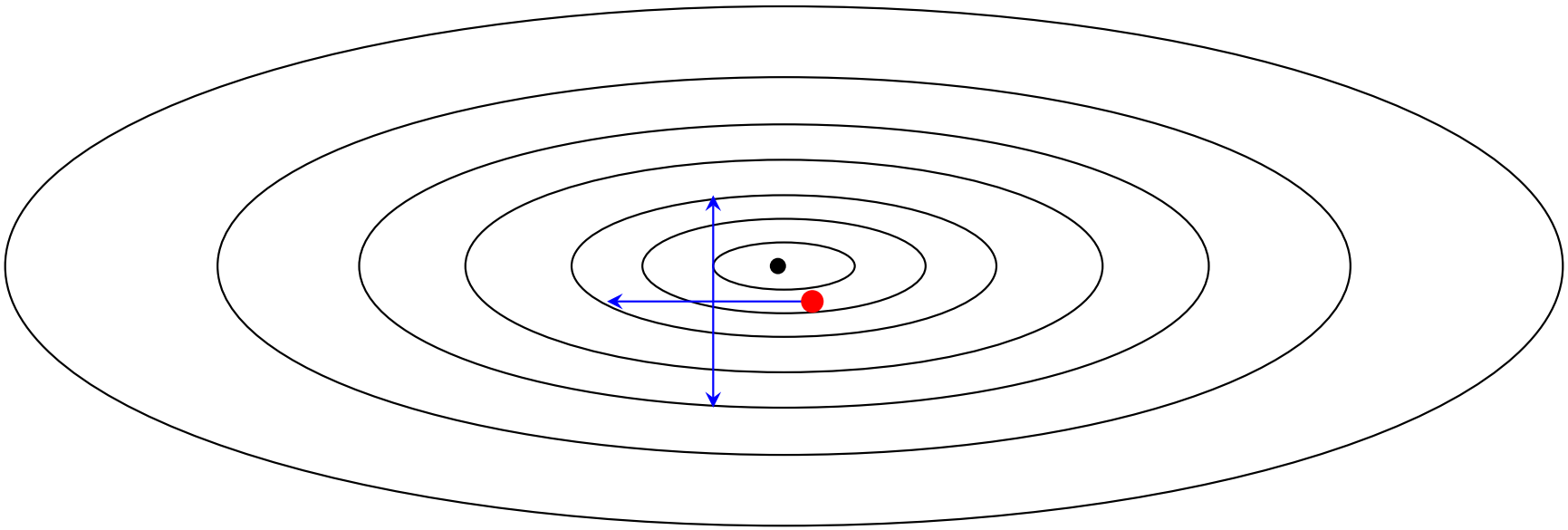
Coordinate search



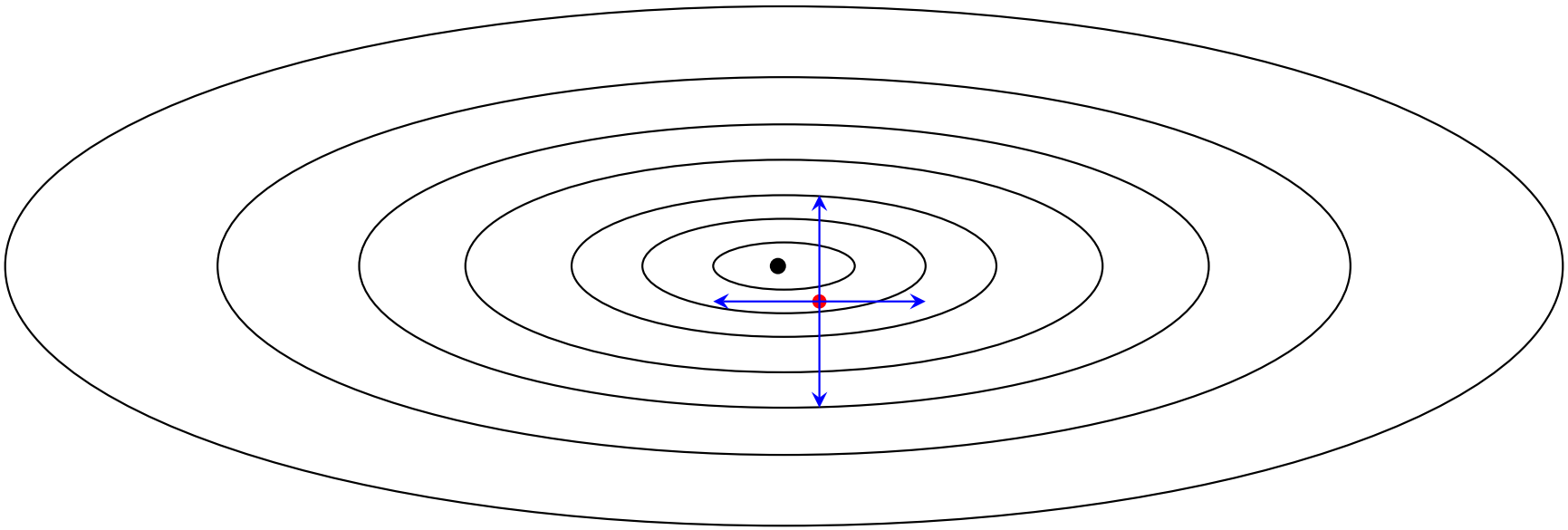
Coordinate search



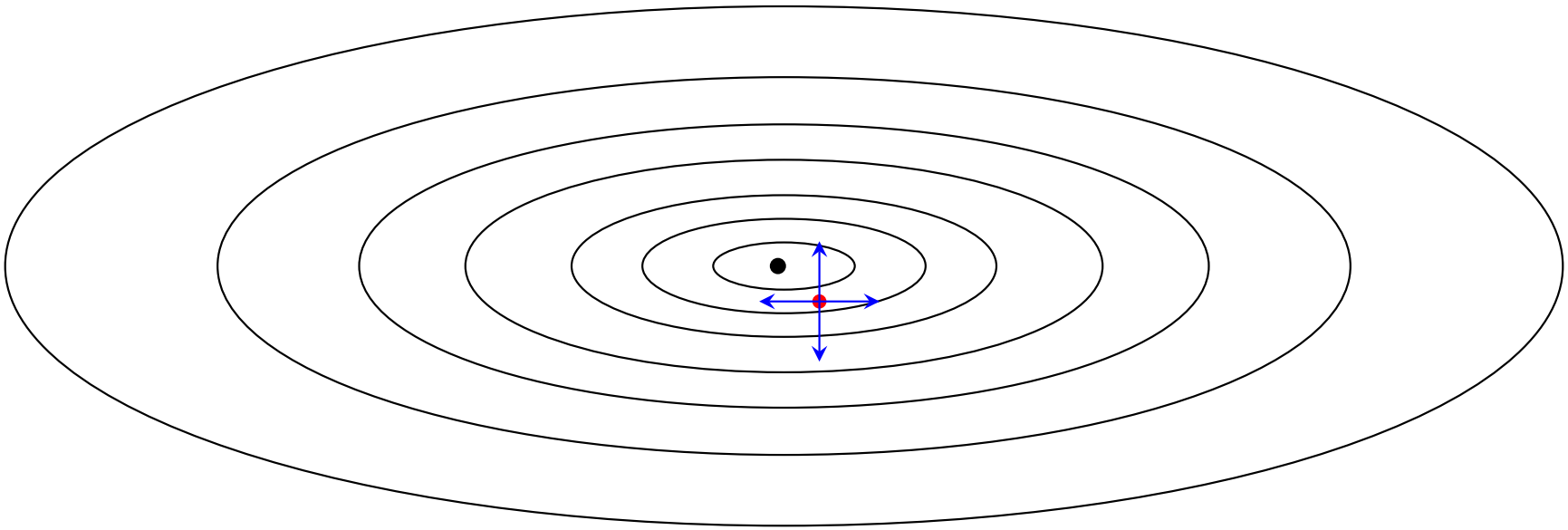
Coordinate search



Coordinate search



Coordinate search



A class of DS methods

Choose: x_0 and α_0.

For $k = 0, 1, 2, \ldots$ (Until α_k is suff. small)

- **Search step (optional)**

A class of DS methods

Choose: x_0 and α_0.

For $k = 0, 1, 2, \ldots$ (Until α_k is suff. small)

- **Search step (optional)**
- **Poll step:** Select D_k PSS and find $x_k + \alpha_k d_k$ $(d_k \in D_k)$:

$$f(x_k + \alpha_k d_k) < f(x_k) - \rho(\alpha_k) \qquad \text{like } \rho(\alpha) = \alpha^2/2.$$

A class of DS methods

Choose: x_0 and α_0.

For $k = 0, 1, 2, \ldots$ (Until α_k is suff. small)

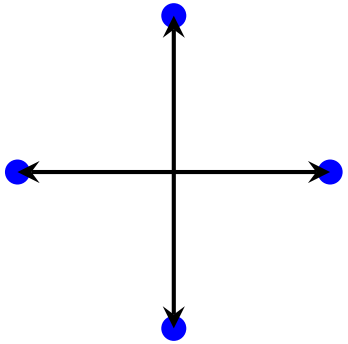
- **Search step (optional)**
- **Poll step:** Select D_k PSS and find $x_k + \alpha_k d_k$ $(d_k \in D_k)$:

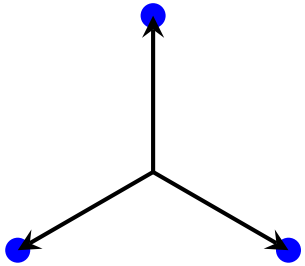
$$f(x_k + \alpha_k d_k) < f(x_k) - \rho(\alpha_k) \qquad \text{like } \rho(\alpha) = \alpha^2/2.$$

- **SUCCESS:** Move $x_{k+1} = x_k + \alpha_k d_k$ and possibly increase $\alpha_{k+1} = \gamma\, \alpha_k$ $(\gamma = 1 \text{ or } 2)$.
- **UNSUCCESS:** Stay $x_{k+1} = x_k$ and decrease $\alpha_{k+1} = \theta\, \alpha_k$ $(\theta = 1/2)$.

Deterministic approach

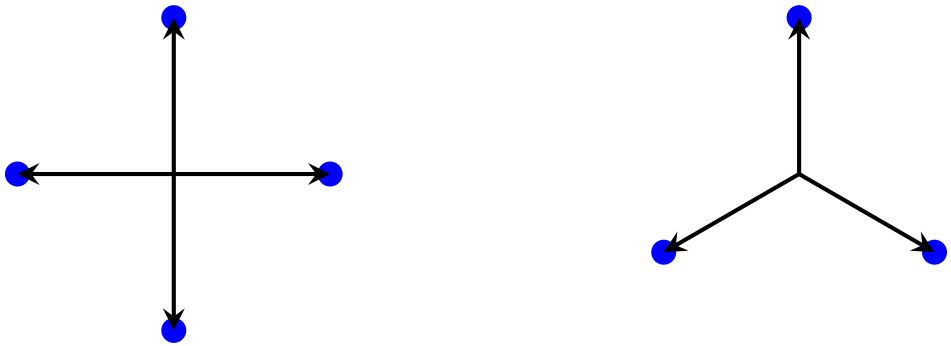
- Positive spanning set (PSS)





Deterministic approach

- Positive spanning set (PSS)

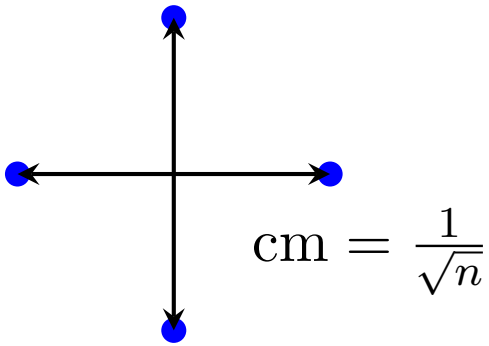


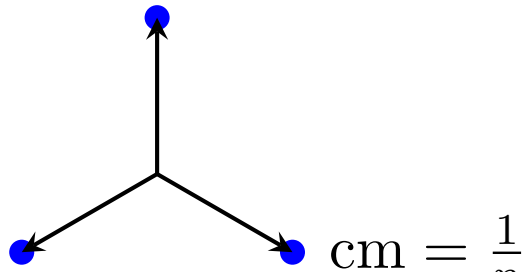
- Cosine measure of a PSS D

$$\mathrm{cm}(D) = \min_{0 \neq v \in \mathbb{R}^n} \max_{d \in D} \frac{d^\top v}{\|d\|\|v\|} > 0.$$

Deterministic approach

- Positive spanning set (PSS)



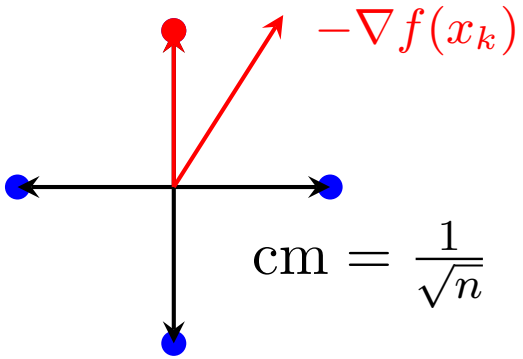


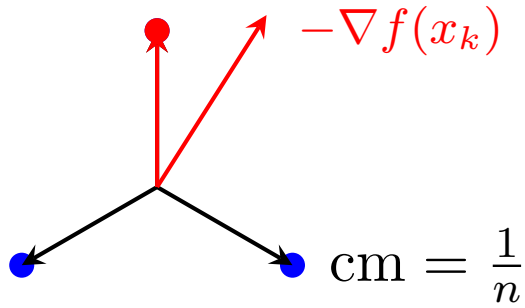
- Cosine measure of a PSS D

$$\mathrm{cm}(D) = \min_{0 \neq v \in \mathbb{R}^n} \max_{d \in D} \frac{d^\top v}{\|d\|\|v\|} > 0.$$

Deterministic approach

- Positive spanning set (PSS)





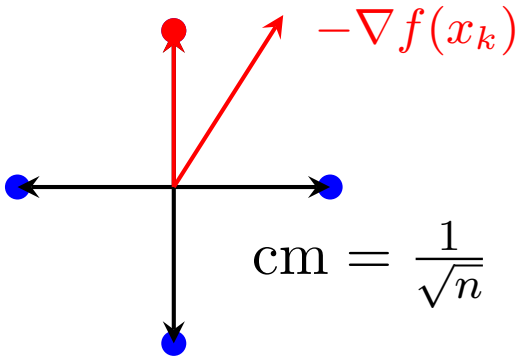
- Cosine measure of a PSS D

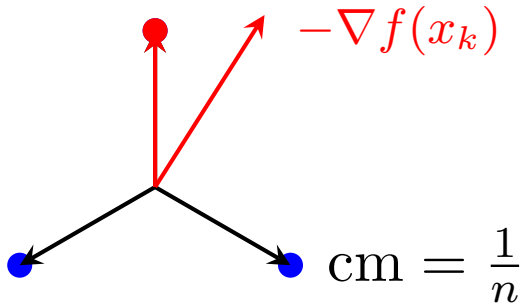
$$\mathrm{cm}(D) = \min_{0 \neq v \in \mathbb{R}^n} \max_{d \in D} \frac{d^\top v}{\|d\|\|v\|} > 0.$$

- Thus $\exists\, d \in D$ descent when $\nabla f(x_k) \neq 0$.

Deterministic approach

- Positive spanning set (PSS)





- Cosine measure of a PSS D

$$\mathrm{cm}(D) = \min_{0 \neq v \in \mathbb{R}^n} \max_{d \in D} \frac{d^\top v}{\|d\|\|v\|} > 0.$$

- Thus $\exists\, d \in D$ descent when $\nabla f(x_k) \neq 0$.

  $\Longrightarrow$ α_k small leads to success!

WCC of DS (smooth case)

Insight: $\underbrace{(\text{decrease in } f) \geq \mathcal{O}(\alpha_k^2)}_{\text{success}} \geq \cdots \geq \underbrace{\mathcal{O}(\alpha_{k_u}^2) \geq \mathcal{O}(\|\nabla f(x_{k_u})\|^2)}_{\text{unsuccess}}$

Kolda, Lewis, Torczon, 2003 SIREV

WCC of DS (smooth case)

Insight: $\underbrace{(\text{decrease in } f) \;\geq\; \mathcal{O}(\alpha_k^2)}_{\text{success}} \;\geq\; \cdots \;\geq\; \underbrace{\mathcal{O}(\alpha_{k_u}^2) \;\geq\; \mathcal{O}(\|\nabla f(x_{k_u})\|^2)}_{\text{unsuccess}}$

Kolda, Lewis, Torczon, 2003 SIREV

Theorem (LNV, 2013 EURO J. Comp. Optim.)

Any such DS method generates a sequence $\{x_k\}_{k\geq 0}$ *such that:*

$$\min_{0\leq j\leq k} \|\nabla f(x_j)\| \;=\; \mathcal{O}(1/\sqrt{k})$$

WCC of DS (smooth case)

Insight: $\underbrace{(\text{decrease in } f) \;\geq\; \mathcal{O}(\alpha_k^2)}_{\text{success}} \;\geq\; \cdots \;\geq\; \underbrace{\mathcal{O}(\alpha_{k_u}^2) \;\geq\; \mathcal{O}(\|\nabla f(x_{k_u})\|^2)}_{\text{unsuccess}}$

Kolda, Lewis, Torczon, 2003 SIREV

Theorem (LNV, 2013 EURO J. Comp. Optim.)

Any such DS method generates a sequence $\{x_k\}_{k\geq 0}$ *such that:*

$$\min_{0\leq j\leq k} \|\nabla f(x_j)\| \;=\; \mathcal{O}(1/\sqrt{k})$$

and takes at most

$$k_\epsilon \;\leq\; \mathcal{O}\left(n\,\epsilon^{-2}\right)$$

iterations to reduce the gradient below $\epsilon \in (0,1)$.

WCC of DS (smooth case)

Insight: $\underbrace{(\text{decrease in } f) \;\geq\; \mathcal{O}(\alpha_k^2)}_{\text{success}} \;\geq\; \cdots \;\geq\; \underbrace{\mathcal{O}(\alpha_{k_u}^2) \;\geq\; \mathcal{O}(\|\nabla f(x_{k_u})\|^2)}_{\text{unsuccess}}$

Kolda, Lewis, Torczon, 2003 SIREV

Theorem (LNV, 2013 EURO J. Comp. Optim.)

Any such DS method generates a sequence $\{x_k\}_{k\geq 0}$ *such that:*

$$\min_{0\leq j\leq k} \|\nabla f(x_j)\| \;=\; \mathcal{O}(1/\sqrt{k})$$

and takes at most

$$k_\epsilon \;\leq\; \mathcal{O}\left(n\,\epsilon^{-2}\right)$$

iterations to reduce the gradient below $\epsilon \in (0,1)$.

- The # of fevals must be multiplied by n: $\mathcal{O}\left(n^2\,\epsilon^{-2}\right)$.

WCC of DS (smooth case)

Insight: $\underbrace{(\text{decrease in } f) \;\geq\; \mathcal{O}(\alpha_k^2)}_{\text{success}} \;\geq\; \cdots \;\geq\; \underbrace{\mathcal{O}(\alpha_{k_u}^2) \;\geq\; \mathcal{O}(\|\nabla f(x_{k_u})\|^2)}_{\text{unsuccess}}$

Kolda, Lewis, Torczon, 2003 SIREV

Theorem (LNV, 2013 EURO J. Comp. Optim.)

Any such DS method generates a sequence $\{x_k\}_{k\geq 0}$ *such that:*

$$\min_{0\leq j\leq k} \|\nabla f(x_j)\| \;=\; \mathcal{O}(1/\sqrt{k})$$

and takes at most

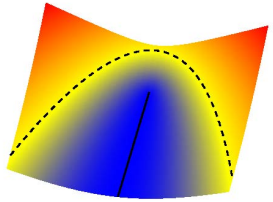
$$k_\epsilon \;\leq\; \mathcal{O}\left(n\,\epsilon^{-2}\right)$$

iterations to reduce the gradient below $\epsilon \in (0,1)$.

- The # of fevals must be multiplied by n: $\mathcal{O}\left(n^2\,\epsilon^{-2}\right)$.
- Bounds depend on $L^2_{\nabla f}$ (instead of $L_{\nabla f}$ as in gradient method).

WCC of DS (smooth, convex case)

Ruling out cases where the supreme distance from the initial level set $L_f(x_0)$ to the solution set X_*^f is infinite...



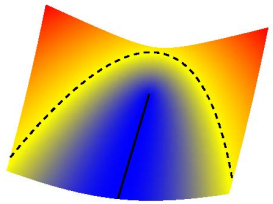
Theorem (M. Dodangeh and LNV, 2016 Math. Program.)

Any such DS method generates a sequence $\{x_k\}_{k\geq 0}$ *such that:*

$$f(x_k) - f_* \;=\; \mathcal{O}(1/k) \qquad k_\epsilon \;\leq\; \mathcal{O}\left(n\,\epsilon^{-1}\right).$$

Again, the # of fevals must be multiplied by n*:* $\mathcal{O}\left(n^2\,\epsilon^{-1}\right)$.

WCC of DS (smooth, convex case)



Ruling out cases where the supreme distance from the initial level set $L_f(x_0)$ to the solution set X_*^f is infinite...

Theorem (M. Dodangeh and LNV, 2016 Math. Program.)

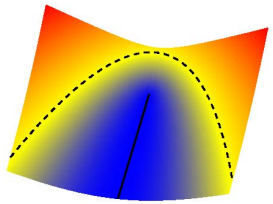
Any such DS method generates a sequence $\{x_k\}_{k\geq 0}$ *such that:*

$$f(x_k) - f_* \;=\; \mathcal{O}(1/k) \qquad k_\epsilon \;\leq\; \mathcal{O}\left(n\,\epsilon^{-1}\right).$$

Again, the # of fevals must be multiplied by n*:* $\mathcal{O}\left(n^2\,\epsilon^{-1}\right)$.

The n^2 factor comes from $\frac{|D|}{\mathrm{cm}(D)^2}$.

WCC of DS (smooth, convex case)



Ruling out cases where the supreme distance from the initial level set $L_f(x_0)$ to the solution set X_*^f is infinite...

Theorem (M. Dodangeh and LNV, 2016 Math. Program.)

Any such DS method generates a sequence $\{x_k\}_{k\geq 0}$ *such that:*

$$f(x_k) - f_* \;=\; \mathcal{O}(1/k) \qquad k_\epsilon \;\leq\; \mathcal{O}\left(n\,\epsilon^{-1}\right).$$

Again, the # of fevals must be multiplied by n*:* $\mathcal{O}\left(n^2\,\epsilon^{-1}\right)$.

The n^2 factor comes from $\frac{|D|}{\mathrm{cm}(D)^2}$. For $D = D_\oplus$ one obtains

$$\frac{2n}{(1/\sqrt{n})^2} \;=\; 2n^2. \qquad \text{Is this optimal?}$$

Optimality of the n^2 factor

Theorem (M. Dodangeh, LNV, and Z. Zhang, 2016 Optim. Lett.)

The factor n^2 is optimal since any PSS D in $\mathbb{R}^n$ satisfies

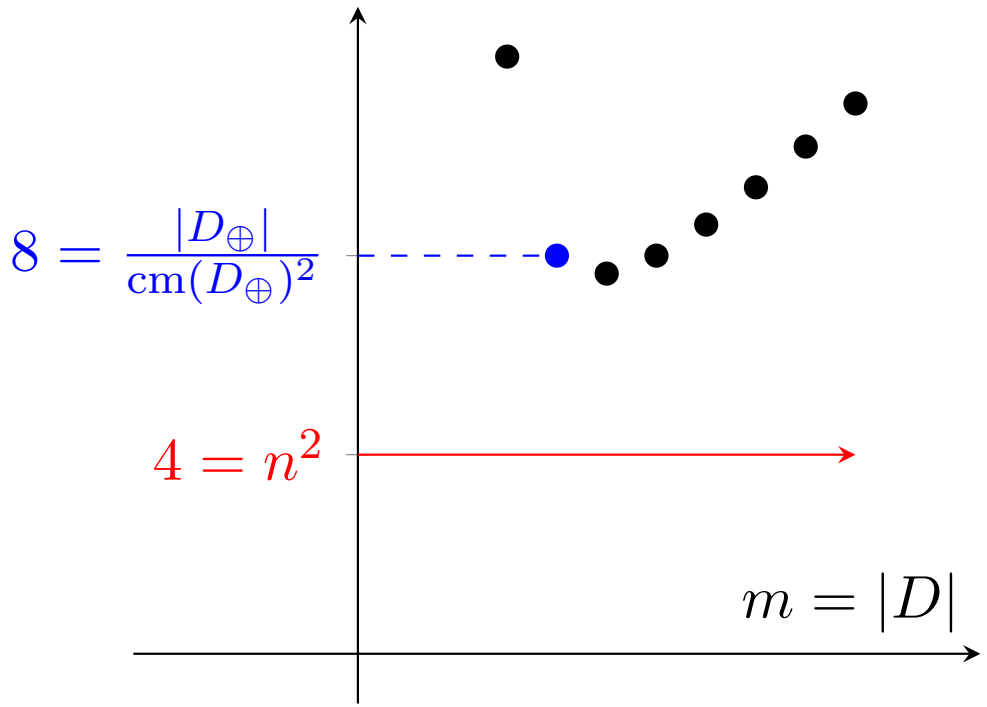
$$\frac{|D|}{\mathrm{cm}(D)^2} \geq \frac{1}{\zeta^2} n^2$$

Optimality of the n^2 factor

Theorem (M. Dodangeh, LNV, and Z. Zhang, 2016 Optim. Lett.)

The factor n^2 is optimal since any PSS D in $\mathbb{R}^n$ satisfies

$$\frac{|D|}{\mathrm{cm}(D)^2} \geq \frac{1}{\zeta^2} n^2$$



Plot of

$$\frac{|D|}{\mathrm{cm}(D)^2} \geq n^2.$$

for the case $n = 2$ and D's with uniform angles,

Global rate of DS (smooth, strongly convex case)

Theorem (M. Dodangeh and LNV, 2014 Math. Program.)

Any such DS method generates a sequence $\{x_k\}_{k\geq 0}$ *such that:*

$$f(x_k) - f_* \ < \ Cr^k,$$

where $r \in (0,1)$ *and* $C > 0$.

Global rate of DS (smooth, strongly convex case)

Theorem (M. Dodangeh and LNV, 2014 Math. Program.)

Any such DS method generates a sequence $\{x_k\}_{k\geq 0}$ *such that:*

$$f(x_k) - f_* \ < \ Cr^k,$$

where $r \in (0,1)$ *and* $C > 0$.

- When f is SC (constant $\mu > 0$), one has (k_0 first unsucc.)

$$\|x_k - x_*\| \ \leq \ \sqrt{L_{\nabla f}/\mu}\, \|x_{k_0} - x_*\|.$$

Global rate of DS (smooth, strongly convex case)

Theorem (M. Dodangeh and LNV, 2014 Math. Program.)

Any such DS method generates a sequence $\{x_k\}_{k\geq 0}$ *such that:*

$$f(x_k) - f_* \ < \ Cr^k,$$

where $r \in (0,1)$ *and* $C > 0$.

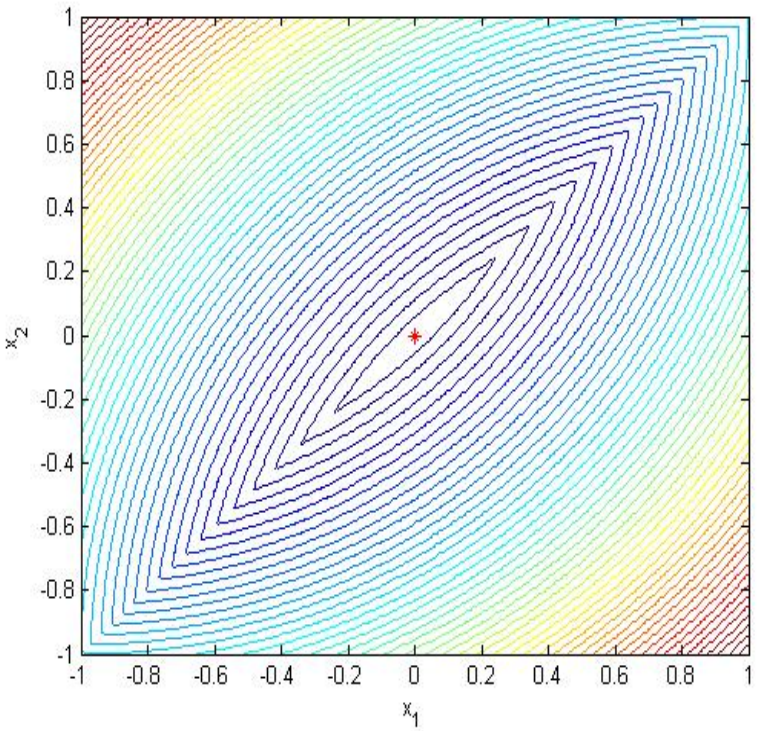
- When f is SC (constant $\mu > 0$), one has (k_0 first unsucc.)

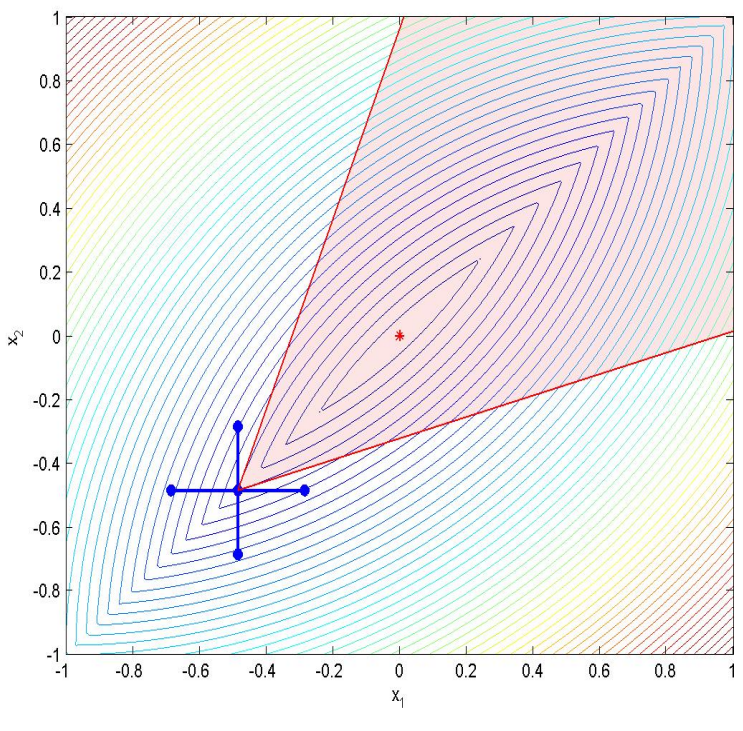
$$\|x_k - x_*\| \ \leq \ \sqrt{L_{\nabla f}/\mu}\, \|x_{k_0} - x_*\|.$$

- A linear rate for the iterates can be derived from

$$\frac{1}{2}\mu\|x - x_*\|^2 \ \leq \ f(x) - f_*.$$

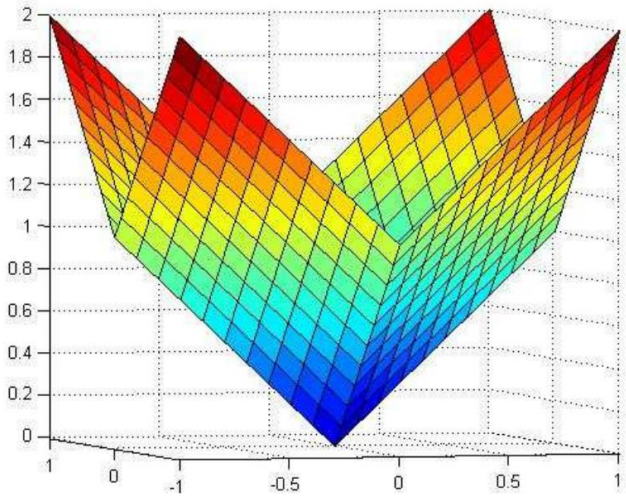
Difficulties in the nonsmooth case



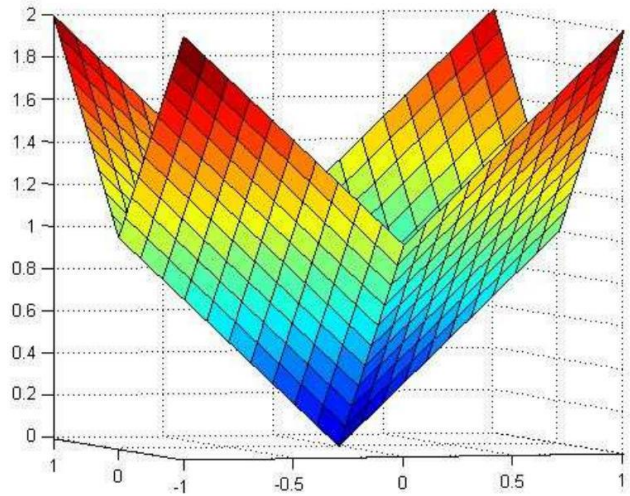


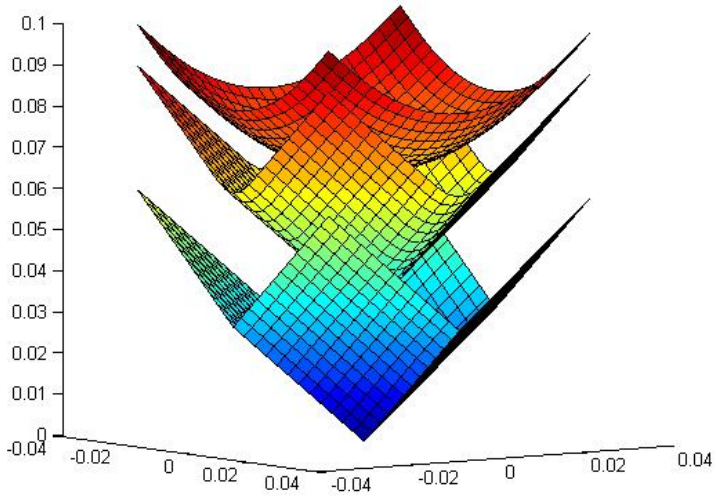
The cone of descent directions at the poll center is shaded.

One possible fix: Combine DS with Smoothing

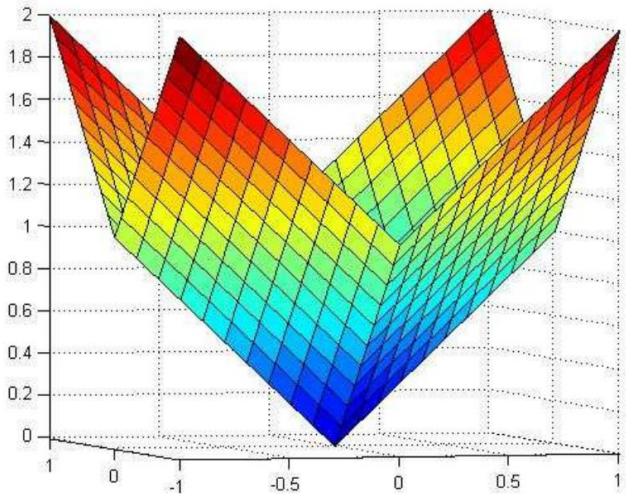


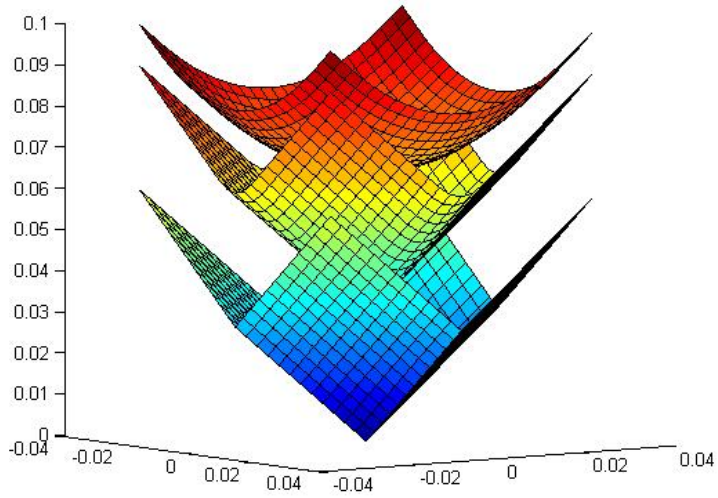
One possible fix: Combine DS with Smoothing





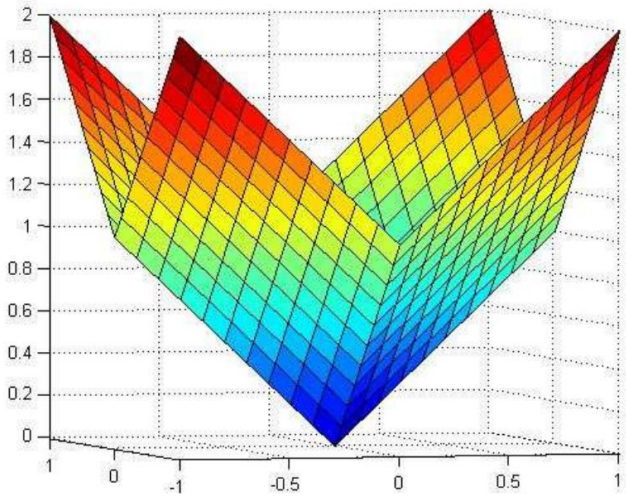
One possible fix: Combine DS with Smoothing

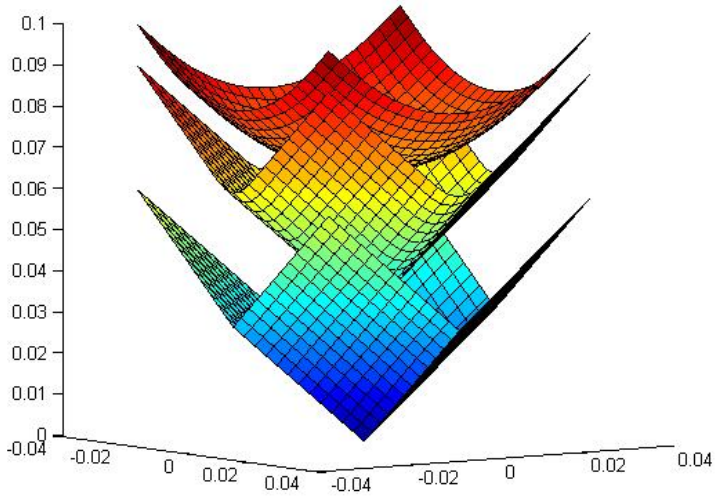




- Essentially the WCC cost increases from $\mathcal{O}\left(\epsilon^{-2}\right)$ to $\mathcal{O}\left(\epsilon^{-3}\right)$ [R. Garmanjani and LNV, 2013 IMA J. Numer. Anal.].

One possible fix: Combine DS with Smoothing





- Essentially the WCC cost increases from $\mathcal{O}\left(\epsilon^{-2}\right)$ to $\mathcal{O}\left(\epsilon^{-3}\right)$ [R. Garmanjani and LNV, 2013 IMA J. Numer. Anal.].
- The # of function evaluations increases from $\mathcal{O}\left(n^2\epsilon^{-2}\right)$ to $\mathcal{O}\left(n^{\frac{5}{2}}\epsilon^{-3}\right)$.

Summary of DS global rates

Imposing sufficient decrease to accept new iterates:

Summary of DS global rates

Imposing sufficient decrease to accept new iterates:

- $\mathcal{O}(-\log(\epsilon))$ strongly convex — linear global rate for f, ∇f, and absolute error in iterates.

Summary of DS global rates

Imposing sufficient decrease to accept new iterates:

- $\mathcal{O}(-\log(\epsilon))$ strongly convex — linear global rate for f, ∇f, and absolute error in iterates.

- $\mathcal{O}(\epsilon^{-1})$ convex — global rate $1/k$ for f and ∇f.

Summary of DS global rates

Imposing sufficient decrease to accept new iterates:

- $\mathcal{O}(-\log(\epsilon))$ strongly convex — linear global rate for f, ∇f, and absolute error in iterates.
- $\mathcal{O}(\epsilon^{-1})$ convex — global rate $1/k$ for f and ∇f.
- $\mathcal{O}(\epsilon^{-2})$ non-convex — global rate $1/\sqrt{k}$ for ∇f.

Summary of DS global rates

Imposing sufficient decrease to accept new iterates:

- $\mathcal{O}(-\log(\epsilon))$ strongly convex — linear global rate for f, ∇f, and absolute error in iterates.
- $\mathcal{O}(\epsilon^{-1})$ convex — global rate $1/k$ for f and ∇f.
- $\mathcal{O}(\epsilon^{-2})$ non-convex — global rate $1/\sqrt{k}$ for ∇f.
- In terms of function evaluations: $\mathcal{O}(n^2\epsilon^{-1})$, $\mathcal{O}(n^2\epsilon^{-2})$. The factor n^2 is proved approximately optimal.

Summary of DS global rates

Imposing sufficient decrease to accept new iterates:

- $\mathcal{O}(-\log(\epsilon))$ strongly convex — linear global rate for f, ∇f, and absolute error in iterates.
- $\mathcal{O}(\epsilon^{-1})$ convex — global rate $1/k$ for f and ∇f.
- $\mathcal{O}(\epsilon^{-2})$ non-convex — global rate $1/\sqrt{k}$ for ∇f.
- In terms of function evaluations: $\mathcal{O}(n^2\epsilon^{-1})$, $\mathcal{O}(n^2\epsilon^{-2})$. The factor n^2 is proved approximately optimal.
- $\mathcal{O}(\epsilon^{-3})$ non-smooth, non-convex — (using smoothing techniques).

Descent condition and successful iterations

Assume the polling directions are normalized.

Lemma

If

$$\operatorname{cm}\left(D_k, -\nabla f(x_k)\right) \geq \kappa \quad \textit{and} \quad \alpha_k < \frac{2\kappa\|\nabla f(x_k)\|}{L_{\nabla f}+1},$$

the k-th iteration is successful.

Descent condition and successful iterations

Assume the polling directions are normalized.

Lemma

If

$$\operatorname{cm}\left(D_k, -\nabla f(x_k)\right) \geq \kappa \quad \textit{and} \quad \alpha_k < \frac{2\kappa\|\nabla f(x_k)\|}{L_{\nabla f}+1},$$

the k-th iteration is successful.

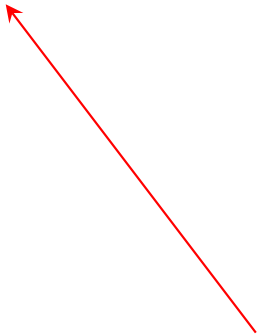
where $\operatorname{cm}(D,v)$ is the cosine measure of D given v, defined by

$$\operatorname{cm}(D,v) = \max_{d\in D} \frac{d^\top v}{\|d\|\|v\|}$$

and $L_{\nabla f}$ is a Lipschitz constant of ∇f.

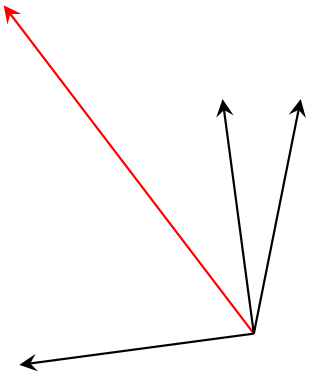
Randomly generating ‘positive spanning sets’ ...

$-\nabla f(x_k)$



Randomly generating ‘positive spanning sets’ ...

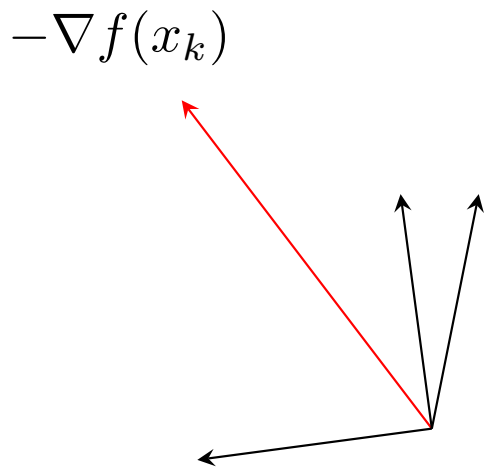
$-\nabla f(x_k)$



$n+1$ random polling directions

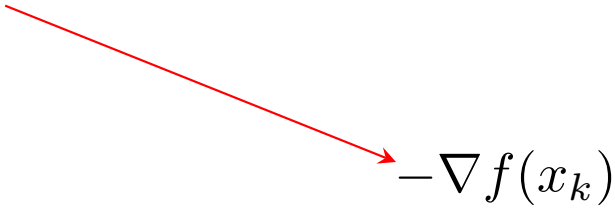
in this case not a PSS

Randomly generating 'positive spanning sets' ...

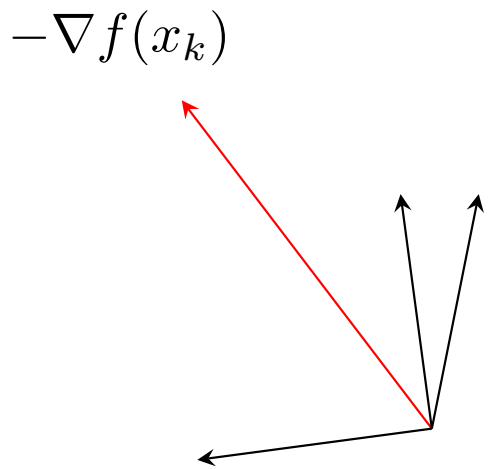


$n+1$ random polling directions

in this case not a PSS

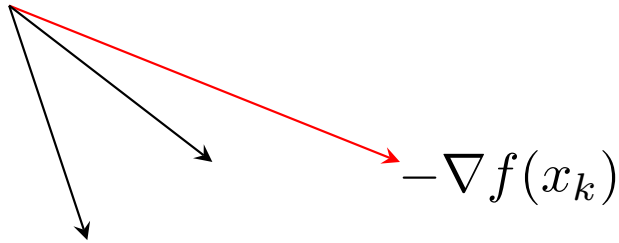


Randomly generating 'positive spanning sets' ...



$n+1$ random polling directions

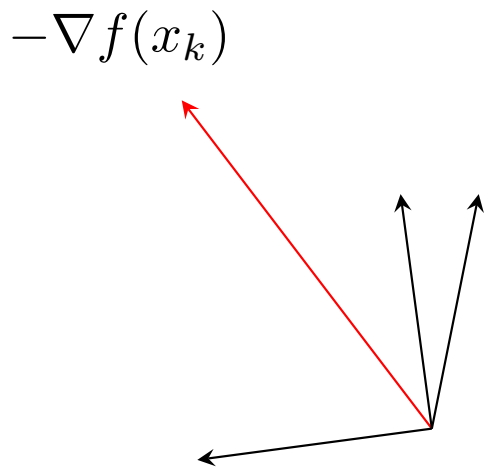
in this case not a PSS



$\leq n$ random polling directions

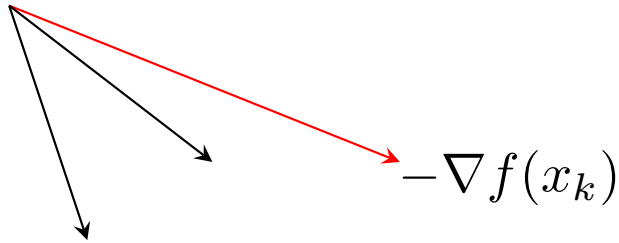
certainly not a PSS ...

Randomly generating 'positive spanning sets' ...



$n+1$ random polling directions

in this case not a PSS



$\leq n$ random polling directions

certainly not a PSS ...

$\mathrm{cm}\big(D_k, -\nabla f(x_k)\big) \geq \kappa$ can be satisfied 'probabilistically' ...

Numerical illustration

Relative performance for different sets of polling directions ($n = 40$).

	$[I \ -I]$	$[Q \ -Q]$	$2n$	$n+1$	$n/4$	2	1
arglina	3.42	8.44	10.30	6.01	1.88	1.00	–
arglinb	20.50	10.35	7.38	2.81	1.85	1.00	2.04
broydn3d	4.33	6.55	6.54	3.59	1.28	1.00	–
dqrtic	7.16	9.37	9.10	4.56	1.70	1.00	–
engval1	10.53	20.89	11.90	6.48	2.08	1.00	2.08
freuroth	56.00	6.33	1.00	1.67	1.67	1.00	4.00
integreq	16.04	16.29	12.44	6.76	2.04	1.00	–
nondquar	6.90	30.23	7.56	4.23	1.87	1.00	–
sinquad	–	–	1.65	2.01	1.00	1.55	–
vardim	1.00	3.80	1.80	2.40	1.80	1.80	4.30

Solution accuracy was 10^{-3}. Averages were taken over 10 independent runs.

Probabilistic descent

From the definition of probabilistic models (Bandeira, LNV, Scheinberg, 2014 SIOPT):

Definition

The sequence $\{\mathfrak{D}_k\}$ *is* (p, κ)*-probabilistically descent if, for each* $k \geq 0$,

$$\Pr\left(\mathrm{cm}(\mathfrak{D}_k, -\nabla f(X_k)) \geq \kappa \mid \mathfrak{D}_0, \ldots, \mathfrak{D}_{k-1}\right) \geq p,$$

Probabilistic descent

From the definition of probabilistic models (Bandeira, LNV, Scheinberg, 2014 SIOPT):

Definition

The sequence $\{\mathfrak{D}_k\}$ *is* (p, κ)*-probabilistically descent if, for each* $k \geq 0$,

$$\Pr\left(\mathrm{cm}(\mathfrak{D}_k, -\nabla f(X_k)) \;\geq\; \kappa \mid \mathfrak{D}_0, \ldots, \mathfrak{D}_{k-1}\right) \geq \; p,$$

Let Z_k be the indicator function of $\left\{\mathrm{cm}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) \geq \kappa\right\}$, and

$$p_0 \;=\; \frac{\ln \theta}{\ln(\gamma^{-1}\theta)} \;=\; \frac{1}{2} \qquad \theta = 1/2,\ \gamma = 2.$$

Global rate: Counting descent

For each realization of the DS algorithm, define

- $\tilde{g}_k$: the gradient with minimum norm among $\nabla f(x_0), \ldots, \nabla f(x_k)$,

Global rate: Counting descent

For each realization of the DS algorithm, define

- $\tilde{g}_k$: the gradient with minimum norm among $\nabla f(x_0), \ldots, \nabla f(x_k)$,
- k_ϵ: the smallest integer such that $\|\nabla f(x_k)\| \leq \epsilon$.

Global rate: Counting descent

For each realization of the DS algorithm, define

- $\tilde{g}_k$: the gradient with minimum norm among $\nabla f(x_0), \ldots, \nabla f(x_k)$,
- k_ϵ: the smallest integer such that $\|\nabla f(x_k)\| \leq \epsilon$.

Denote the corresponding random variables by $\tilde{G}_k$ and K_ϵ.

Global rate: Counting descent

For each realization of the DS algorithm, define

- $\tilde{g}_k$: the gradient with minimum norm among $\nabla f(x_0), \ldots, \nabla f(x_k)$,
- k_ϵ: the smallest integer such that $\|\nabla f(x_k)\| \leq \epsilon$.

Denote the corresponding random variables by $\tilde{G}_k$ and K_ϵ.

Let z_ℓ denote the realization of $Z_\ell = \left\{\mathrm{cm}\big(\mathfrak{D}_\ell, -\nabla f(X_\ell)\big) \geq \kappa\right\}$ $(\ell \geq 0)$.

Intuition: If $\|\tilde{g}_k\|$ is 'big', then $\sum_{\ell=0}^{k-1} z_\ell$ is probably 'small'.

Global rate: Counting descent

In fact, one can prove

$$\sum_{\ell=0}^{k-1} z_\ell \;\leq\; \mathcal{O}\left(\frac{1}{\|\tilde{g}_k\|^2}\right) + p_0 k.$$

Global rate: Counting descent

In fact, one can prove

$$\sum_{\ell=0}^{k-1} z_\ell \;\leq\; \mathcal{O}\left(\frac{1}{\|\tilde{g}_k\|^2}\right) + p_0 k.$$

It then results,

$$\left\{\|\tilde{G}_k\| > \epsilon\right\} \;\subset\; \left\{\sum_{\ell=0}^{k-1} Z_\ell \leq \left[\mathcal{O}\left(\frac{1}{k\epsilon^2}\right) + p_0\right] k\right\}.$$

Global rate: Counting descent

In fact, one can prove

$$\sum_{\ell=0}^{k-1} z_\ell \;\leq\; \mathcal{O}\left(\frac{1}{\|\tilde{g}_k\|^2}\right) + p_0 k.$$

It then results,

$$\left\{\|\tilde{G}_k\| > \epsilon\right\} \;\subset\; \left\{\sum_{\ell=0}^{k-1} Z_\ell \leq \underbrace{\left[\mathcal{O}\left(\frac{1}{k\epsilon^2}\right) + p_0\right]}_{\lambda} k\right\}.$$

Global rate: Counting descent

In fact, one can prove

$$\sum_{\ell=0}^{k-1} z_\ell \;\leq\; \mathcal{O}\left(\frac{1}{\|\tilde{g}_k\|^2}\right) + p_0 k.$$

It then results,

$$\left\{\|\tilde{G}_k\| > \epsilon\right\} \;\subset\; \left\{\sum_{\ell=0}^{k-1} Z_\ell \leq \underbrace{\left[\mathcal{O}\left(\frac{1}{k\epsilon^2}\right) + p_0\right]}_{\lambda} k\right\}.$$

Hence

$$\Pr\left(\|\tilde{G}_k\| \leq \epsilon\right) \;=\; 1 - \Pr\left(\|\tilde{G}_k\| > \epsilon\right) \;\geq\; 1 - \underbrace{\Pr\left(\sum_{\ell=0}^{k-1} Z_\ell \leq \lambda\, k\right)}_{\text{apply Chernoff \& Submartingale Theory}}.$$

Global rate and WCC bound

Theorem (Gratton, Royer, LNV, and Zhang, 2015 SIOPT)

Suppose that $\{\mathfrak{D}_k\}$ *is* (p, κ)*-probabilistically descent with* $p > p_0$*. Then*

$$\Pr\left(\|\tilde{G}_k\| \leq \mathcal{O}\left(\frac{1}{\kappa\sqrt{k}}\right)\right) \geq 1 - \exp\left[-\mathcal{O}(k)\right].$$

Global rate and WCC bound

Theorem (Gratton, Royer, LNV, and Zhang, 2015 SIOPT)

Suppose that $\{\mathfrak{D}_k\}$ *is* (p, κ)*-probabilistically descent with* $p > p_0$*. Then*

$$\Pr\left(\|\tilde{G}_k\| \leq \mathcal{O}\left(\frac{1}{\kappa\sqrt{k}}\right)\right) \geq 1 - \exp\left[-\mathcal{O}(k)\right].$$

$\longrightarrow$ $\mathcal{O}(1/\sqrt{k})$ sublinear rate with overwhelmingly high probability.

Global rate and WCC bound

Theorem (Gratton, Royer, LNV, and Zhang, 2015 SIOPT)

Suppose that $\{\mathfrak{D}_k\}$ *is* (p, κ)*-probabilistically descent with* $p > p_0$*. Then*

$$\Pr\left(\|\tilde{G}_k\| \leq \mathcal{O}\left(\frac{1}{\kappa\sqrt{k}}\right)\right) \geq 1 - \exp\left[-\mathcal{O}(k)\right].$$

$\longrightarrow$ $\mathcal{O}(1/\sqrt{k})$ sublinear rate with overwhelmingly high probability.

Since $\Pr(K_\epsilon \leq k) = \Pr(\|\tilde{G}_k\| \leq \epsilon)$, we also get:

Theorem (Gratton, Royer, LNV, and Zhang, 2015 SIOPT)

Suppose that $\{\mathfrak{D}_k\}$ *is* (p, κ)*-probabilistically descent with* $p > p_0$*. Then*

$$\Pr\left(K_\epsilon \leq \left\lceil \mathcal{O}\left(\frac{\epsilon^{-2}}{\kappa^2}\right)\right\rceil\right) \geq 1 - \exp\left[-\mathcal{O}(\epsilon^{-2})\right].$$

Global rate and WCC bound

Theorem (Gratton, Royer, LNV, and Zhang, 2015 SIOPT)

Suppose that $\{\mathfrak{D}_k\}$ *is* (p, κ)*-probabilistically descent with* $p > p_0$*. Then*

$$\Pr\left(\|\tilde{G}_k\| \leq \mathcal{O}\left(\frac{1}{\kappa\sqrt{k}}\right)\right) \geq 1 - \exp\left[-\mathcal{O}(k)\right].$$

$\longrightarrow$ $\mathcal{O}(1/\sqrt{k})$ sublinear rate with overwhelmingly high probability.

Since $\Pr(K_\epsilon \leq k) = \Pr(\|\tilde{G}_k\| \leq \epsilon)$, we also get:

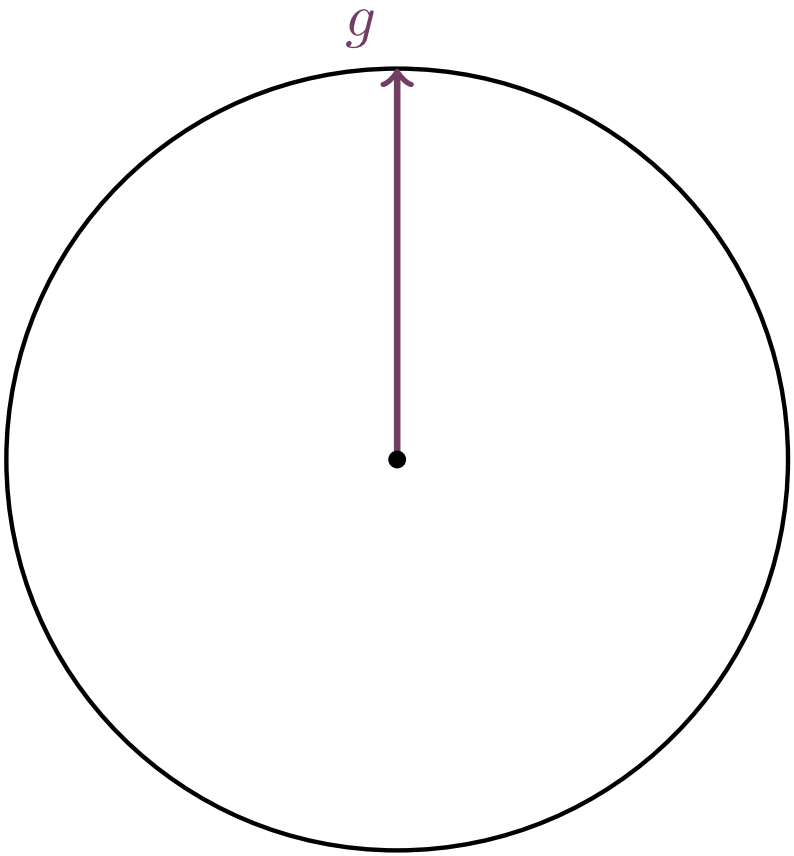
Theorem (Gratton, Royer, LNV, and Zhang, 2015 SIOPT)

Suppose that $\{\mathfrak{D}_k\}$ *is* (p, κ)*-probabilistically descent with* $p > p_0$*. Then*

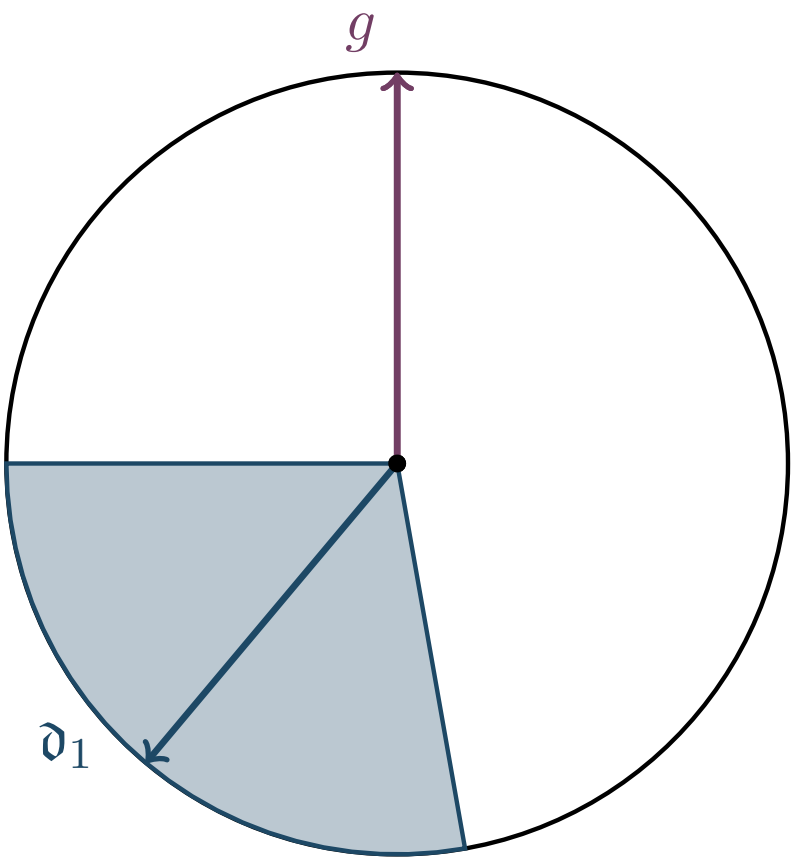
$$\Pr\left(K_\epsilon \leq \left\lceil \mathcal{O}\left(\frac{\epsilon^{-2}}{\kappa^2}\right)\right\rceil\right) \geq 1 - \exp\left[-\mathcal{O}(\epsilon^{-2})\right].$$

$\longrightarrow$ $\mathcal{O}(\epsilon^{-2})$ bound for # of iter. with overwhelmingly high probability.

Two uniform directions are enough, one is not

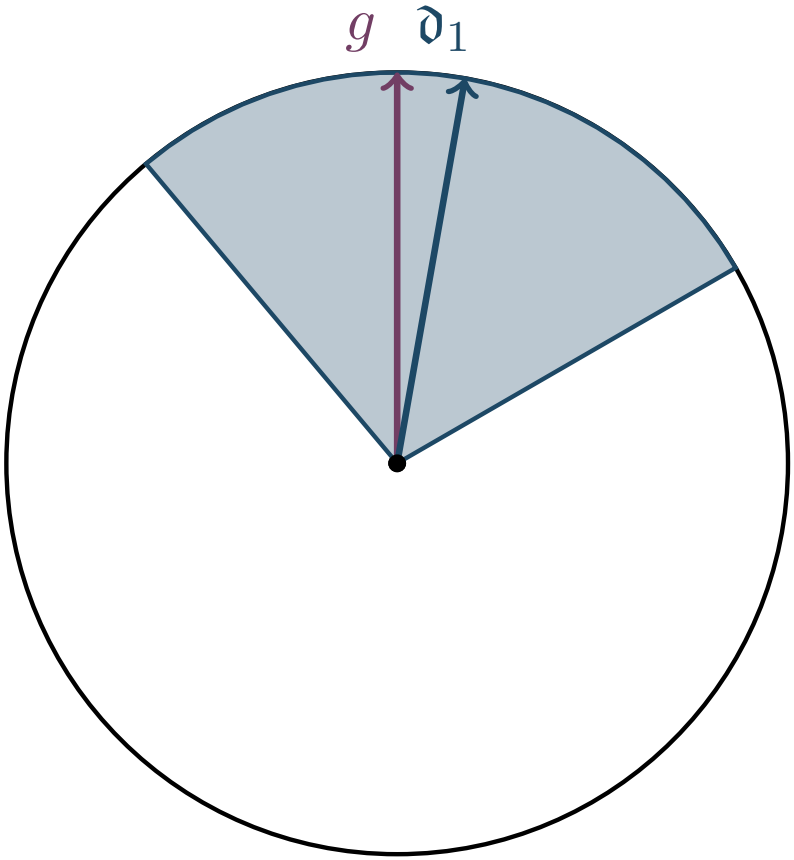


Two uniform directions are enough, one is not



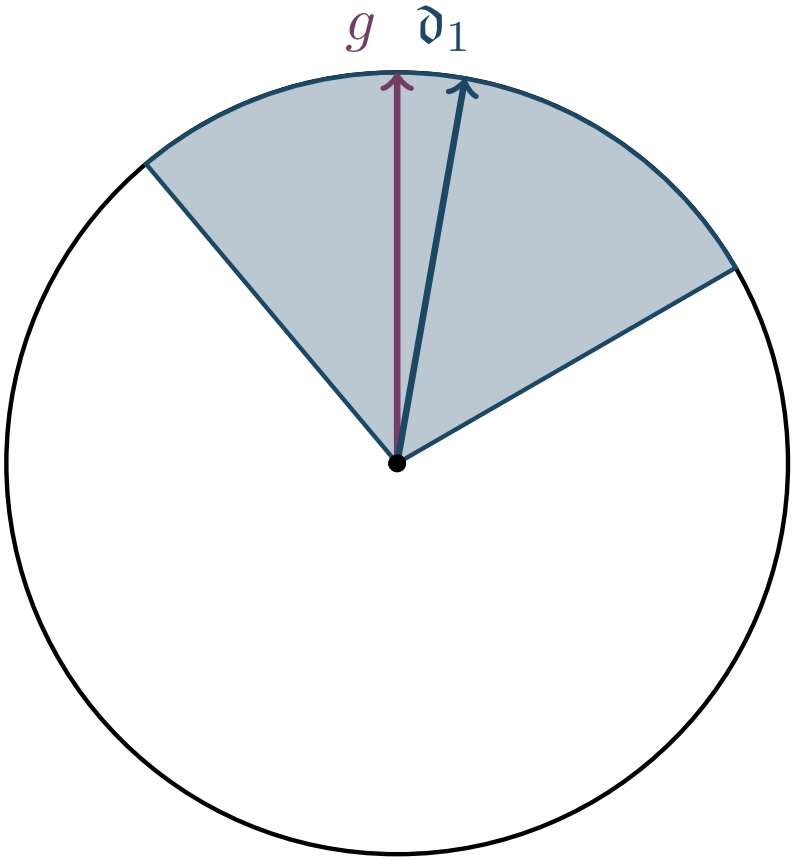
$$\mathfrak{d}_1 \sim \mathcal{U}(\mathbb{S}^1) \Rightarrow \forall \kappa \in (0,1), \quad \Pr\left(\mathrm{cm}\left(\mathfrak{d}_1, g\right) = \mathfrak{d}_1^\top g \geq \kappa\right) < 1/2.$$

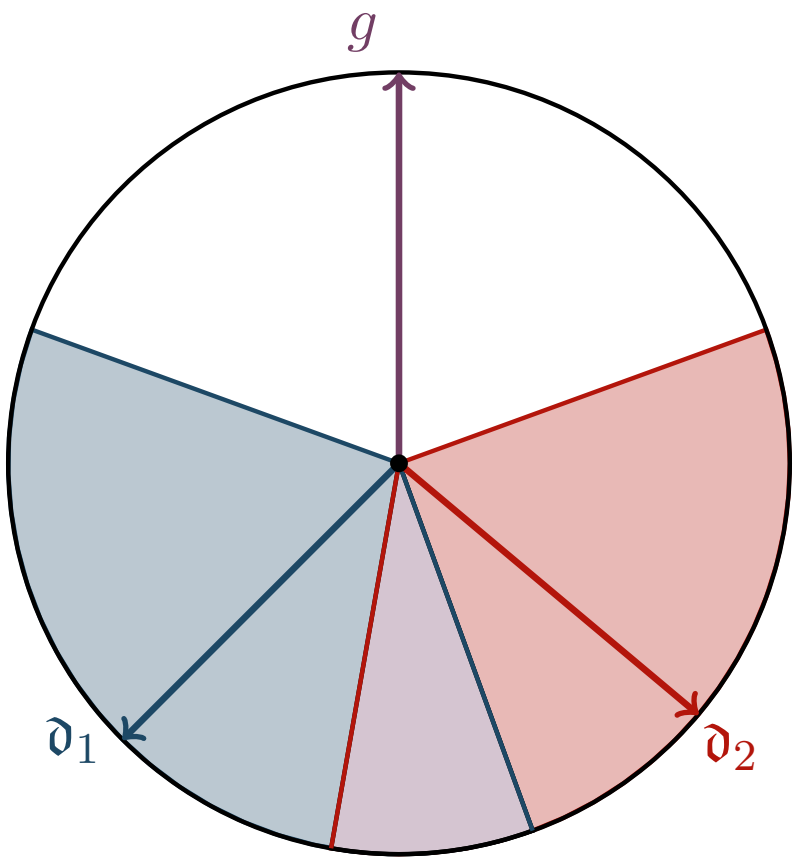
Two uniform directions are enough, one is not



$$\mathfrak{d}_1 \sim \mathcal{U}(\mathbb{S}^1) \Rightarrow \forall \kappa \in (0,1), \quad \Pr\left(\mathrm{cm}\left(\mathfrak{d}_1, g\right) = \mathfrak{d}_1^\top g \geq \kappa\right) < 1/2.$$

Two uniform directions are enough, one is not





$$\mathfrak{d}_1 \sim \mathcal{U}(\mathbb{S}^1) \Rightarrow \forall \kappa \in (0,1), \quad \Pr\left(\mathrm{cm}\left(\mathfrak{d}_1, g\right) = \mathfrak{d}_1^\top g \geq \kappa\right) < 1/2.$$

$$\mathfrak{d}_1, \mathfrak{d}_2 \sim \mathcal{U}(\mathbb{S}^1) \Rightarrow \exists \kappa^* \in (0,1), \quad \Pr\left(\mathrm{cm}\left(\{\mathfrak{d}_1, \mathfrak{d}_2\}, g\right) \geq \kappa^*\right) > 1/2.$$

Worst case complexity: Dependence on the dimension

Then, when $r = |D| > 1$, $\{\mathfrak{D}_k\}$ is p-probabilistically $(1/\sqrt{n})$-descent for some $p > p_0 = 1/2$ independent of n.

Worst case complexity: Dependence on the dimension

Then, when $r = |D| > 1$, $\{\mathfrak{D}_k\}$ is p-probabilistically $(1/\sqrt{n})$-descent for some $p > p_0 = 1/2$ independent of n.

Plugging $\kappa = 1/\sqrt{n}$ into the WCC bound, one obtains

Worst case complexity: Dependence on the dimension

Then, when $r = |D| > 1$, $\{\mathfrak{D}_k\}$ is p-probabilistically $(1/\sqrt{n})$-descent for some $p > p_0 = 1/2$ independent of n.

Plugging $\kappa = 1/\sqrt{n}$ into the WCC bound, one obtains

WCC (number of function evaluations)

$$\Pr\left(K_\epsilon^f \leq \left\lceil \mathcal{O}\left(n\epsilon^{-2}\right)\right\rceil r\right) \geq 1 - \exp\left[-\mathcal{O}(\epsilon^{-2})\right].$$

Worst case complexity: Dependence on the dimension

Then, when $r = |D| > 1$, $\{\mathfrak{D}_k\}$ is p-probabilistically $(1/\sqrt{n})$-descent for some $p > p_0 = 1/2$ independent of n.

Plugging $\kappa = 1/\sqrt{n}$ into the WCC bound, one obtains

WCC (number of function evaluations)

$$\Pr\left(K_\epsilon^f \leq \left\lceil \mathcal{O}\left(n\epsilon^{-2}\right)\right\rceil r\right) \geq 1 - \exp\left[-\mathcal{O}(\epsilon^{-2})\right].$$

The WCC bound is $\mathcal{O}(rn\epsilon^{-2})$, better than when $\mathcal{O}(n^2\epsilon^{-2})$ when $r \ll n$.

Worst case complexity: Dependence on the dimension

Then, when $r = |D| > 1$, $\{\mathfrak{D}_k\}$ is p-probabilistically $(1/\sqrt{n})$-descent for some $p > p_0 = 1/2$ independent of n.

Plugging $\kappa = 1/\sqrt{n}$ into the WCC bound, one obtains

WCC (number of function evaluations)

$$\Pr\left(K_\epsilon^f \leq \left\lceil \mathcal{O}\left(n\epsilon^{-2}\right)\right\rceil r\right) \geq 1 - \exp\left[-\mathcal{O}(\epsilon^{-2})\right].$$

The WCC bound is $\mathcal{O}(rn\epsilon^{-2})$, better than when $\mathcal{O}(n^2\epsilon^{-2})$ when $r \ll n$.

Theory admits $r = 2$, leading to

$$\mathcal{O}(n\epsilon^{-2}) \quad !!!$$

A more detailed look at the numerical experiments

Relative performance for different sets of polling directions ($n = 40$).

	$[I \ -I]$	$[Q \ -Q]$	2 ($\gamma = 2$)	4 ($\gamma = 1.1$)
arglina	1.00	3.17	5.86	6.73
arglinb	34.12	5.34	1.00	2.02
broydn3d	1.00	1.91	2.04	3.47
dqrtic	1.18	1.36	1.00	1.48
engval1	1.05	1.00	2.29	2.89
freuroth	17.74	7.39	1.35	1.00
integreq	1.54	1.49	1.00	1.34
nondquar	1.00	2.82	1.37	1.73
sinquad	–	1.26	1.00	–
vardim	20.31	11.02	1.00	1.84

Now $\gamma = 1$ for $[I \ -I]$ and $[Q \ -Q]$.

A more detailed look at the numerical experiments

Relative performance for different sets of polling directions ($n = 100$).

	$[I \ -I]$	$[Q \ -Q]$	2 ($\gamma = 2$)	4 ($\gamma = 1.1$)
arglina	1.00	3.86	5.86	7.58
arglinb	138.28	107.32	1.00	1.99
broydn3d	1.00	2.57	1.92	3.21
dqrtic	3.01	3.25	1.00	1.46
engval1	1.04	1.00	2.06	2.84
freuroth	31.94	17.72	1.36	1.00
integreq	1.83	1.66	1.00	1.22
nondquar	1.18	2.83	1.00	1.17
sinquad	–	–	–	–
vardim	112.22	19.72	1.00	2.36

Now $\gamma = 1$ for $[I \ -I]$ and $[Q \ -Q]$.

So, a new proof technique

A new proof technique for establishing global rates and WCC bounds for randomized algorithms for which

So, a new proof technique

A new proof technique for establishing global rates and WCC bounds for randomized algorithms for which

- the new iterate depends on some object (directions, models),
- the quality of the object is favorable with a certain probability.

So, a new proof technique

A new proof technique for establishing global rates and WCC bounds for randomized algorithms for which

- the new iterate depends on some object (directions, models),
- the quality of the object is favorable with a certain probability.

The technique is based on:

- counting the number of iterations for which the quality is favorable,
- examining the probabilistic behavior of this number.

So, a new proof technique

A new proof technique for establishing global rates and WCC bounds for randomized algorithms for which

- the new iterate depends on some object (directions, models),
- the quality of the object is favorable with a certain probability.

The technique is based on:

- counting the number of iterations for which the quality is favorable,
- examining the probabilistic behavior of this number.

It is thus possible to obtain a rate of $\mathcal{O}(1/\sqrt{k})$, with overwhelmingly high probability, also for the TRM based on probabilistic models [Gratton, Royer, LNV, and Zhang, 2017].

Two linearly constrained problems

Linear equality constraints

$$\begin{cases} \min_{x \in \mathbb{R}^n} & f(x) \\ \text{s.t.} & Ax = b. \end{cases}$$

- Equivalent to the unconstrained problem $\min_{\tilde{x} \in \mathbb{R}^{n-m}} f\left(x_0 + W\tilde{x}\right)$ with $W \in \mathbb{R}^{n \times (n-m)}$ orthonormal basis for $\mathrm{null}(A)$ and $Ax_0 = b$.
- Deterministic and probabilistic approaches/analyses apply!

Two linearly constrained problems

Linear equality constraints

$$\begin{cases} \min_{x \in \mathbb{R}^n} & f(x) \\ \text{s.t.} & Ax = b. \end{cases}$$

- Equivalent to the unconstrained problem $\min_{\tilde{x} \in \mathbb{R}^{n-m}} f\left(x_0 + W\tilde{x}\right)$ with $W \in \mathbb{R}^{n \times (n-m)}$ orthonormal basis for $\text{null}(A)$ and $Ax_0 = b$.
- Deterministic and probabilistic approaches/analyses apply!

Bounds

$$\begin{cases} \min_{x \in \mathbb{R}^n} & f(x) \\ \text{s.t.} & l \leq x \leq u. \end{cases}$$

- Deterministic practice: Uses $D_{\oplus} = \{e_1, \ldots, e_n, -e_1, \ldots, -e_n\}$ to guarantee convergence and moves parallel to the constraints.

A class of DS methods (linear constraints)

Choose: x_0 and α_0.

For $k = 0, 1, 2, \ldots$ (Until α_k is suff. small)

- **Search step (optional)**

A class of DS methods (linear constraints)

Choose: x_0 and α_0.

For $k = 0, 1, 2, \ldots$ (Until α_k is suff. small)

- **Search step (optional)**
- **Poll step:** Select D_k $(\ldots)$ and find $x_k + \alpha_k d_k$ FEASIBLE $(d_k \in D_k)$:

$$f(x_k + \alpha_k d_k) < f(x_k) - \rho(\alpha_k) \qquad \text{like } \rho(\alpha) = \alpha^2/2.$$

A class of DS methods (linear constraints)

Choose: x_0 and α_0.

For $k = 0, 1, 2, \ldots$ (Until α_k is suff. small)

- **Search step (optional)**
- **Poll step:** Select D_k $(\ldots)$ and find $x_k + \alpha_k d_k$ FEASIBLE $(d_k \in D_k)$:

$$f(x_k + \alpha_k d_k) < f(x_k) - \rho(\alpha_k) \qquad \text{like } \rho(\alpha) = \alpha^2/2.$$

- **SUCCESS:** Move $x_{k+1} = x_k + \alpha_k d_k$ and possibly increase $\alpha_{k+1} = \gamma\, \alpha_k$ $(\gamma = 1 \text{ or } 2)$.
- **UNSUCCESS:** Stay $x_{k+1} = x_k$ and decrease $\alpha_{k+1} = \theta\, \alpha_k$ $(\theta = 1/2)$.

Nearby / approximate active constraints

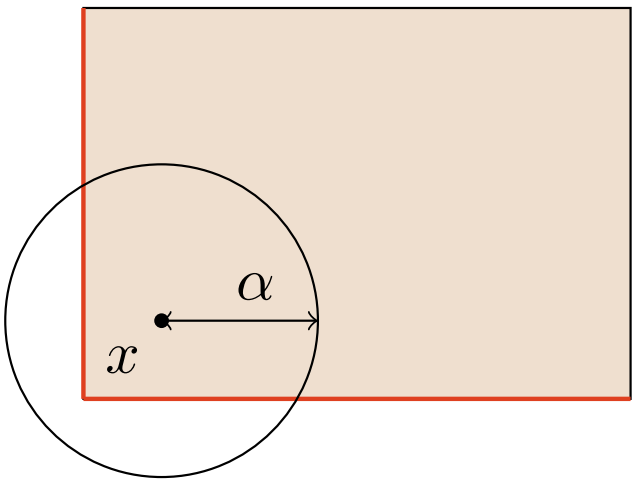
Using the bounds case as example, where the feasible set is $\mathcal{F} = \{l \leq x \leq u\}$, one has

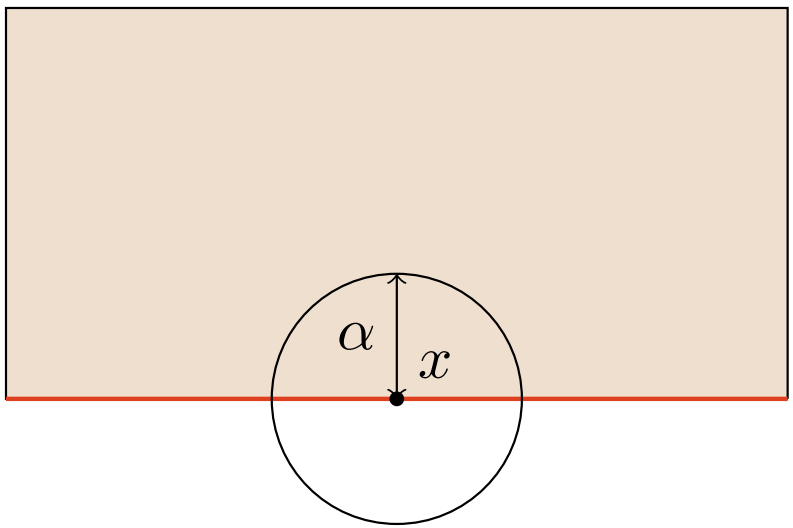
Nearby constraints

The indexes

$$\begin{aligned} I_u(x, \alpha) &= \{i : |u_i - [x]_i| \leq \alpha\} \\ I_l(x, \alpha) &= \{i : |l_i - [x]_i| \leq \alpha\} \end{aligned}$$

define the nearby constraints at $x \in \mathcal{F}$ given $\alpha > 0$.



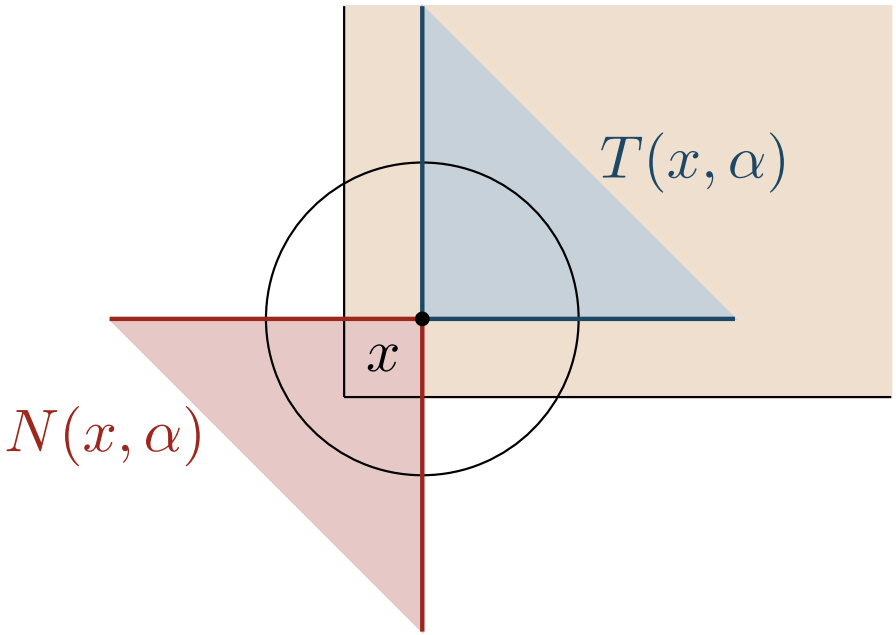


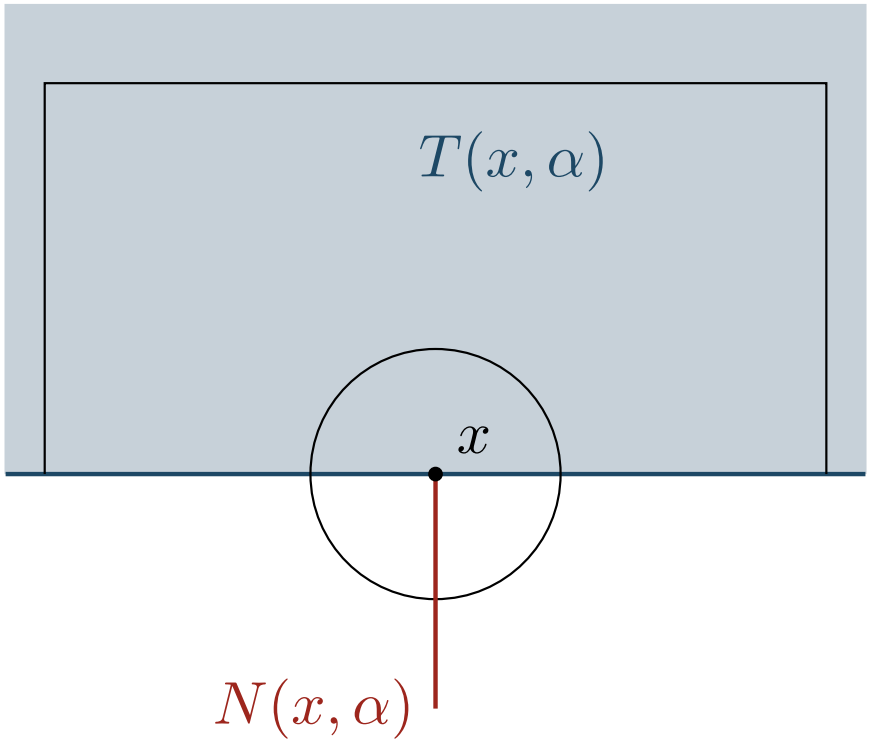
Approximate tangent/normal cones

- **Approximate normal cone $N(x, \alpha)$:** Positive span of

$$\{e_i\}_{i \in I_u(x,\alpha)} \cup \{-e_i\}_{i \in I_l(x,\alpha)} .$$

- **Approximate tangent cone $T(x, \alpha)$:** polar of $N(x, \alpha)$.





Feasible descent property

- Recall the cosine measure that identifies descent sets

$$\mathrm{cm}(D, -\nabla f(x)) \;=\; \max_{d \in D} \frac{d^\top [-\nabla f(x)]}{\|d\| \| -\nabla f(x)\|} \;\geq\; \kappa.$$

Feasible descent property

- Recall the cosine measure that identifies descent sets

$$\mathrm{cm}(D, -\nabla f(x)) \;=\; \max_{d \in D} \frac{d^\top [-\nabla f(x)]}{\|d\| \| -\nabla f(x)\|} \;\geq\; \kappa.$$

Feasible descent property

D is a κ-feasible descent set for $T(x, \alpha)$ if $D \subset T(x, \alpha)$ and

$$\mathrm{cm}_{T(x,\alpha)}(D, -\nabla f(x)) \;=\; \max_{d \in D} \frac{d^\top [-\nabla f(x)]}{\|d\| \|P_{T(x,\alpha)}[-\nabla f(x)]\|} \;\geq\; \kappa.$$

Feasible descent property

- Recall the cosine measure that identifies descent sets

$$\mathrm{cm}(D, -\nabla f(x)) \;=\; \max_{d \in D} \frac{d^\top [-\nabla f(x)]}{\|d\| \| -\nabla f(x)\|} \;\geq\; \kappa.$$

Feasible descent property

D is a κ-feasible descent set for $T(x, \alpha)$ if $D \subset T(x, \alpha)$ and

$$\mathrm{cm}_{T(x,\alpha)}(D, -\nabla f(x)) \;=\; \max_{d \in D} \frac{d^\top [-\nabla f(x)]}{\|d\| \|P_{T(x,\alpha)}[-\nabla f(x)]\|} \;\geq\; \kappa.$$

- Using κ-feasible descent sets guarantees both convergence and complexity (analysis similar to the unconstrained case).
- For bounds only, $D_\oplus \cap T(x, \alpha)$ is always $\frac{1}{\sqrt{n}}$-feasible descent.

Probabilistic feasible descent

Definition

The sequence $\{\mathfrak{D}_k\}$ is (p, κ)-probabilistically feasible descent if, for each $k \geq 0$,

$$\Pr\left(\mathrm{cm}_{T_k}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) \geq \kappa \mid \mathfrak{D}_0, \ldots, \mathfrak{D}_{k-1}\right) \geq p.$$

where $T_k = T(X_k, A_k)$.

Probabilistic feasible descent

Definition

The sequence $\{\mathfrak{D}_k\}$ is (p, κ)-probabilistically feasible descent if, for each $k \geq 0$,

$$\Pr\left(\mathrm{cm}_{T_k}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) \geq \kappa \mid \mathfrak{D}_0, \ldots, \mathfrak{D}_{k-1}\right) \geq p.$$

where $T_k = T(X_k, A_k)$.

Convergence and Complexity

If $\{\mathfrak{D}_k\}$ is (p, κ)-probabilistically feasible descent with $p > p_0$,

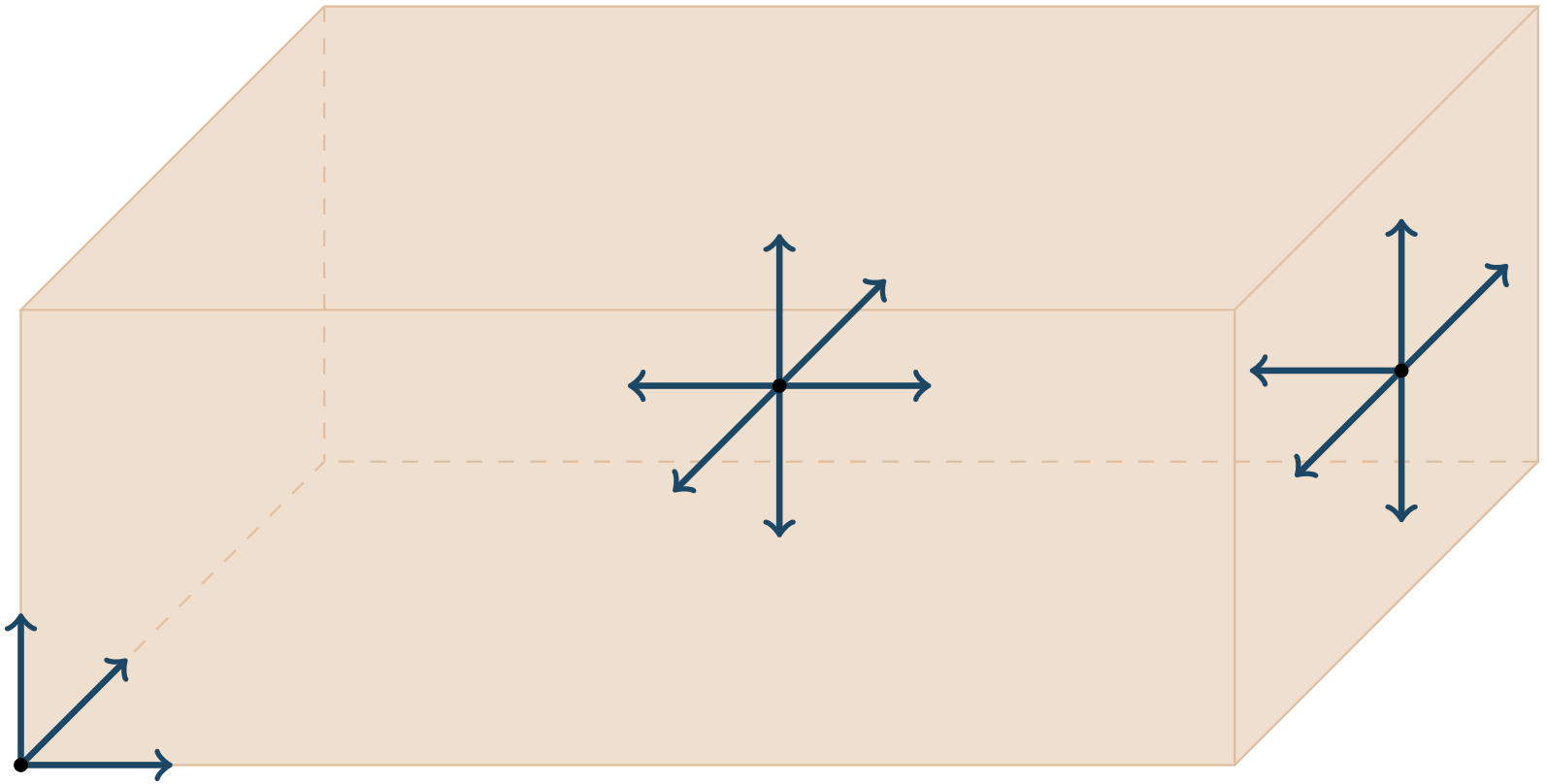
- **almost-sure convergence towards stationary points,**
- **complexity bound for ϵ-stationarity:**

$$\Pr\left(K_\epsilon \leq \left\lceil \mathcal{O}\left(\frac{\epsilon^{-2}}{\kappa^2}\right)\right\rceil\right) \geq 1 - \exp\left[-\mathcal{O}(\epsilon^{-2})\right].$$

Possible direction generation techniques

Random subset of the generators

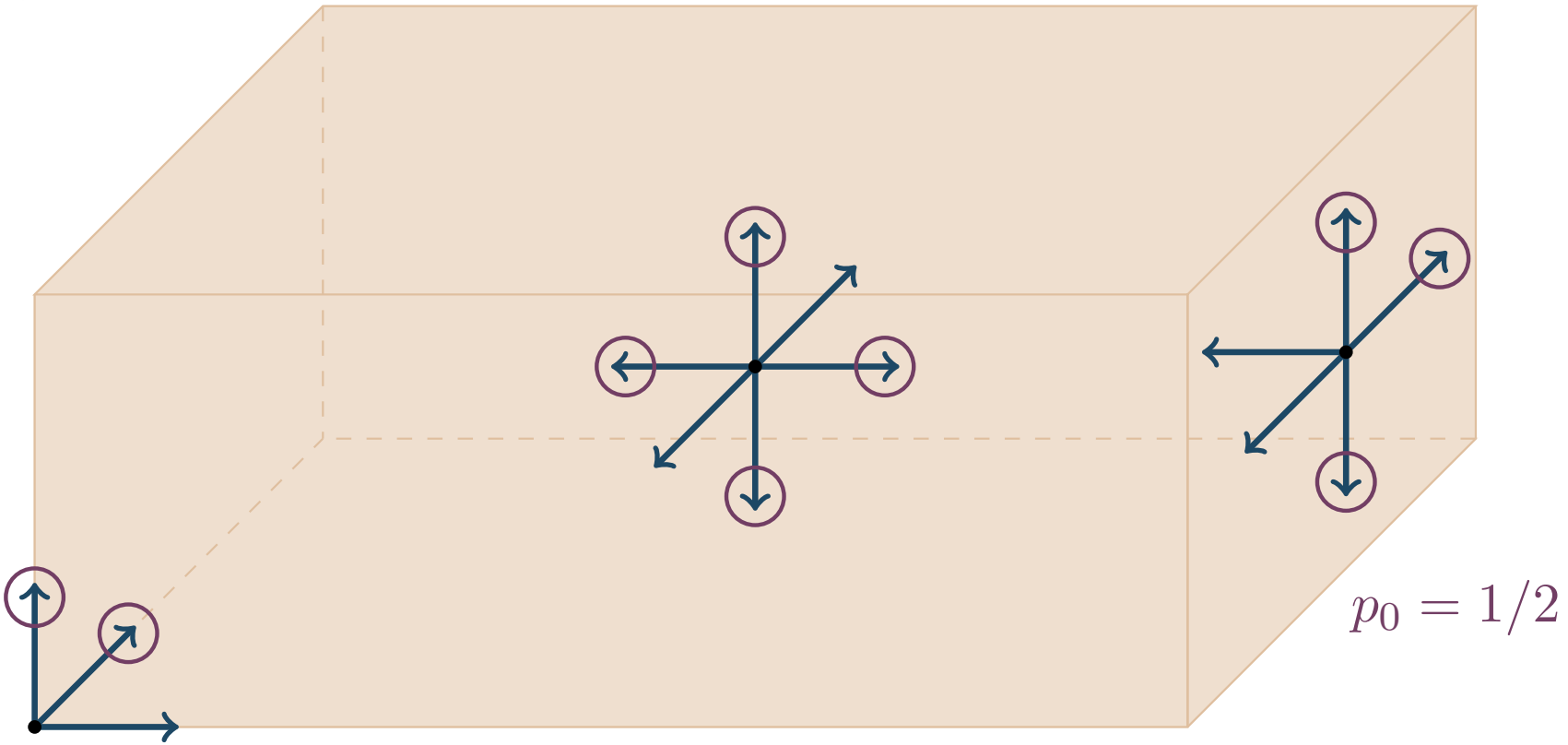
1. Compute a deterministic set of positive generators V_k for T_k.



Possible direction generation techniques

Random subset of the generators

1. Compute a deterministic set of positive generators V_k for T_k.
2. Take a random sample $\mathfrak{D}_k$ of V_k of size $> |V_k| p_0$.
3. Then $\{\mathfrak{D}_k\}$ is (p, κ)-probabilistically feasible descent with $p > p_0$.



Using fewer directions by exploiting subspaces

Idea

- In unconstrained optimization, probabilistic descent can use very few directions... and directions/cones lie in $\mathbb{R}^n$...
- We thus need to identify subspaces in the approximate tangent cones!

Using fewer directions by exploiting subspaces

Idea

- In unconstrained optimization, probabilistic descent can use very few directions... and directions/cones lie in $\mathbb{R}^n$...
- We thus need to identify subspaces in the approximate tangent cones!

Lemma

Let S_k be a linear subspace within a cone T_k. Then

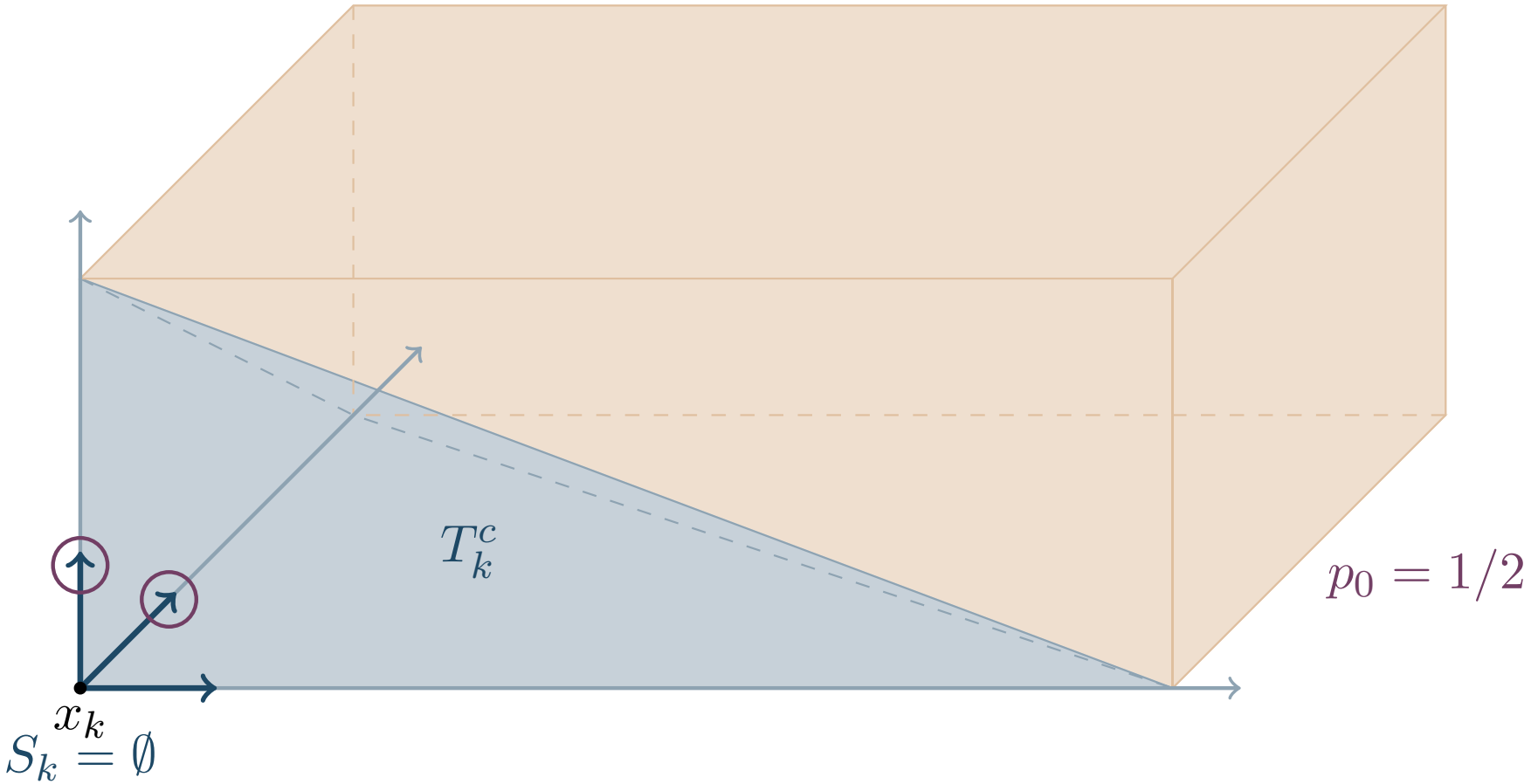
$$T_k \;=\; S_k + T_k^c,$$

where T_k^c is a cone lying in $S_k^{\perp}$.

Using fewer directions by exploiting subspaces (2)

Two types of directions

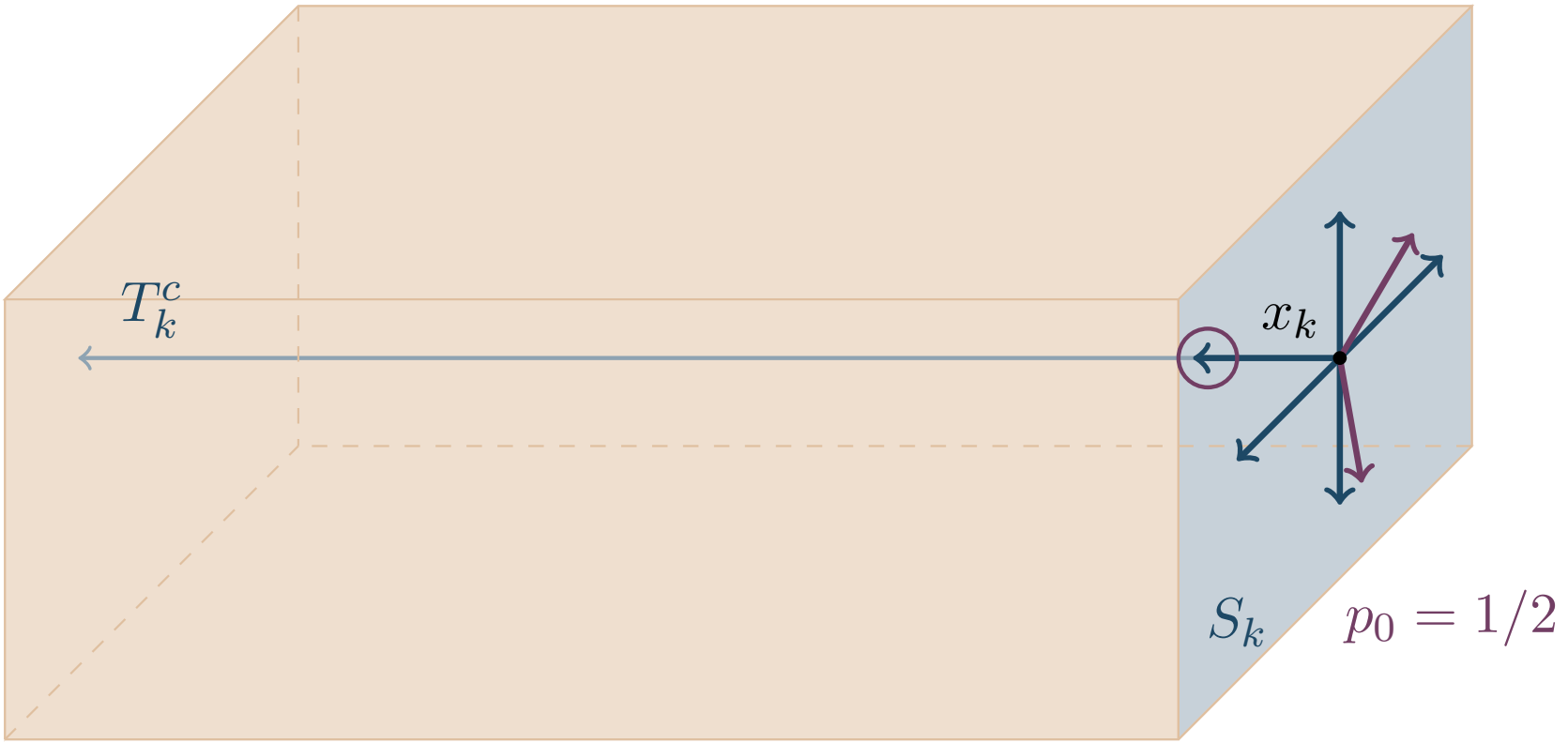
- Subspace S_k: Generate directions randomly.
- Orthogonal part T_k^c: Use a random subset of its positive generators.



Using fewer directions by exploiting subspaces (2)

Two types of directions

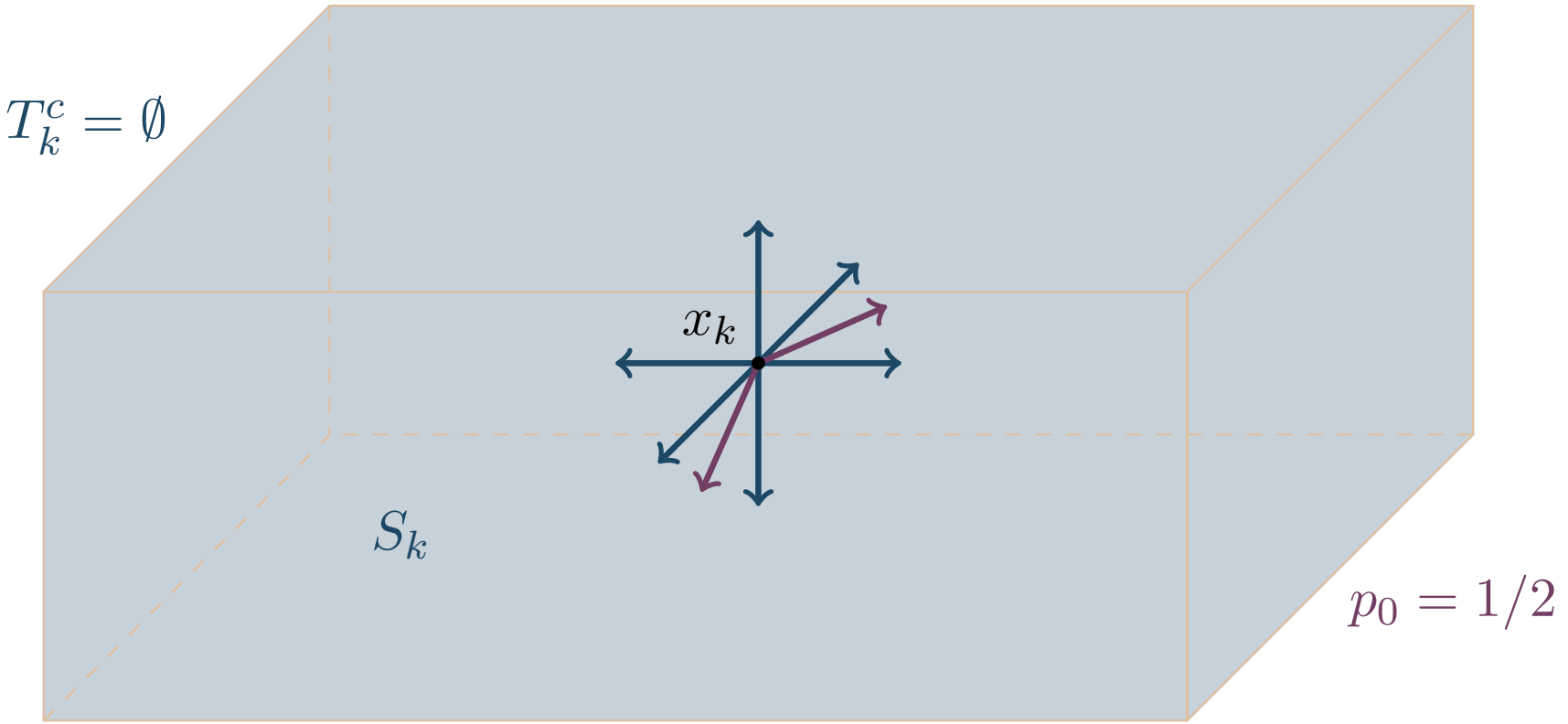
- Subspace S_k: Generate directions randomly.
- Orthogonal part T_k^c: Use a random subset of its positive generators.



Using fewer directions by exploiting subspaces (2)

Two types of directions

- Subspace S_k: Generate directions randomly.
- Orthogonal part T_k^c: Use a random subset of its positive generators.



Complexity (function evaluations)

- The general bound for K_ϵ^f is $\mathcal{O}\left(r\kappa^{-2}\epsilon^{-2}\right)$.

Complexity (function evaluations)

- The general bound for K_ϵ^f is $\mathcal{O}\left(r\kappa^{-2}\epsilon^{-2}\right)$.
- The linear constrained cased is as the unconstrained one (with $n \longleftrightarrow n-m$).

Complexity (function evaluations)

- The general bound for K_ϵ^f is $\mathcal{O}\left(r\kappa^{-2}\epsilon^{-2}\right)$.
- The linear constrained cased is as the unconstrained one (with $n \longleftrightarrow n-m$).
- In the bounds case:

Only $n_b < n$ variables are bounded

Method	r	κ	Bound
Determ.	$2n$	$\frac{1}{\sqrt{n}}$	$\mathcal{O}\left(n^2\epsilon^{-2}\right)$
Proba. 1	$\mathcal{O}(2np_0)$	$\frac{1}{\sqrt{n}}$	$\mathcal{O}\left(n^2\epsilon^{-2}\right)$
Proba. 2 (subspace)	$\mathcal{O}(1) + \mathcal{O}\left(n_b\, p_0\right)$	$\frac{1}{\sqrt{n}}$	$\mathcal{O}\left(n\, n_b\epsilon^{-2}\right)$

Numerical experiments — Bound constraints

- Comparison with MATLAB built-in `patternsearch` function.

Four solvers

Name	Polling in $T(x_k, \alpha_k) = T_k = S_k + T_k^c$	Guarantee
`dspfd-0`	Shuffled $D_\oplus \cap T_k$	Deterministic
`dspfd-1`	Random subset of $D_\oplus \cap T_k$	Probabilistic
`dspfd-2`	(Two) random vectors in S_k Random subset of $D_\oplus \cap T_k^c$	Probabilistic
`patternsearch`	$D_\oplus \cap T_k$	Deterministic

Numerical experiments — Bound constraints

- Comparison with MATLAB built-in `patternsearch` function.

Four solvers

Name	Polling in $T(x_k, \alpha_k) = T_k = S_k + T_k^c$	Guarantee
`dspfd-0`	Shuffled $D_\oplus \cap T_k$	Deterministic
`dspfd-1`	Random subset of $D_\oplus \cap T_k$	Probabilistic
`dspfd-2`	(Two) random vectors in S_k Random subset of $D_\oplus \cap T_k^c$	Probabilistic
`patternsearch`	$D_\oplus \cap T_k$	Deterministic

Performance profiles

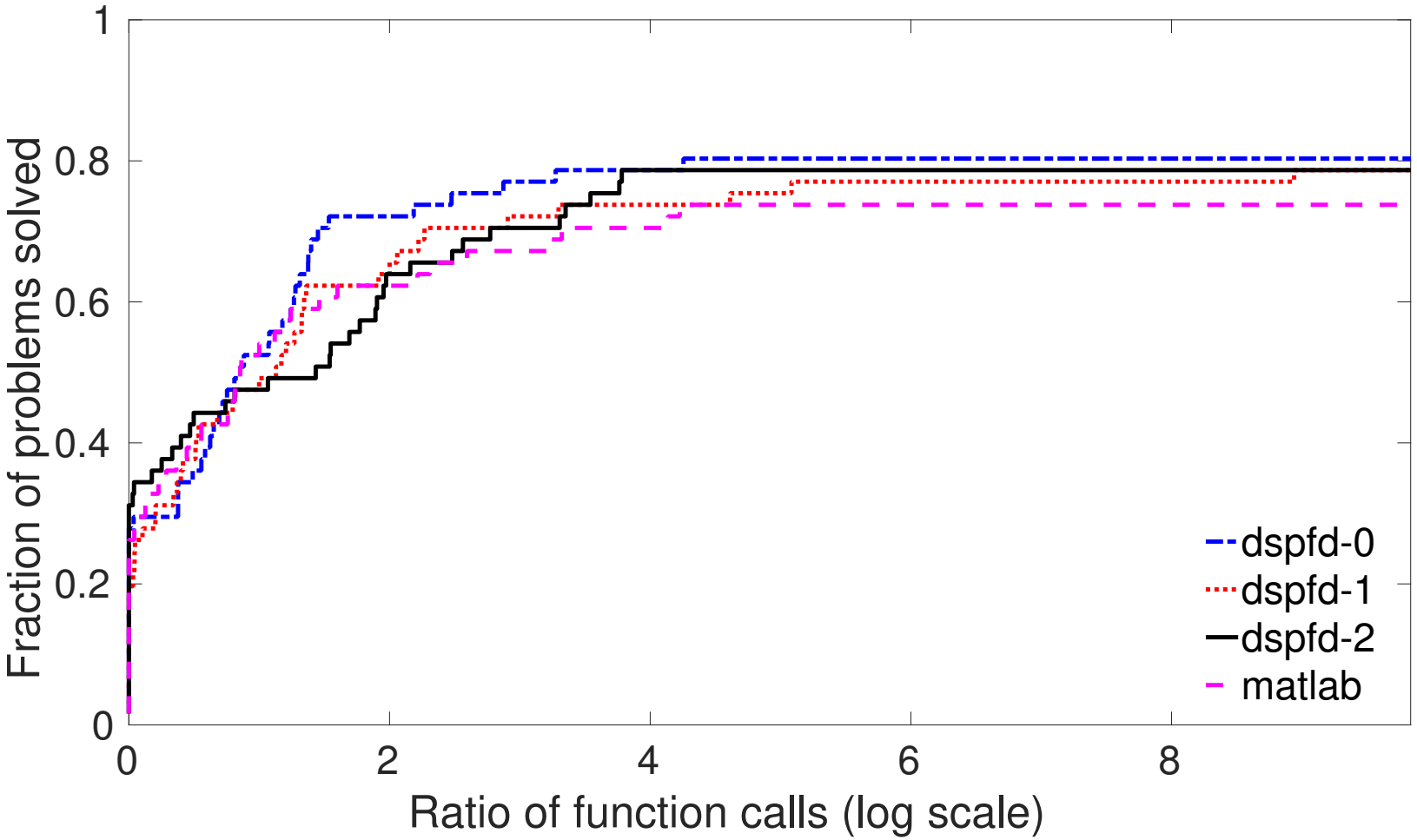
- Criterion: # of function evaluations (budget of $2000n$) to satisfy

$$f(x_k) - f_{best} < 10^{-3}(f(x_0) - f_{best}).$$

- Benchmark: Problems from the CUTEst collection.

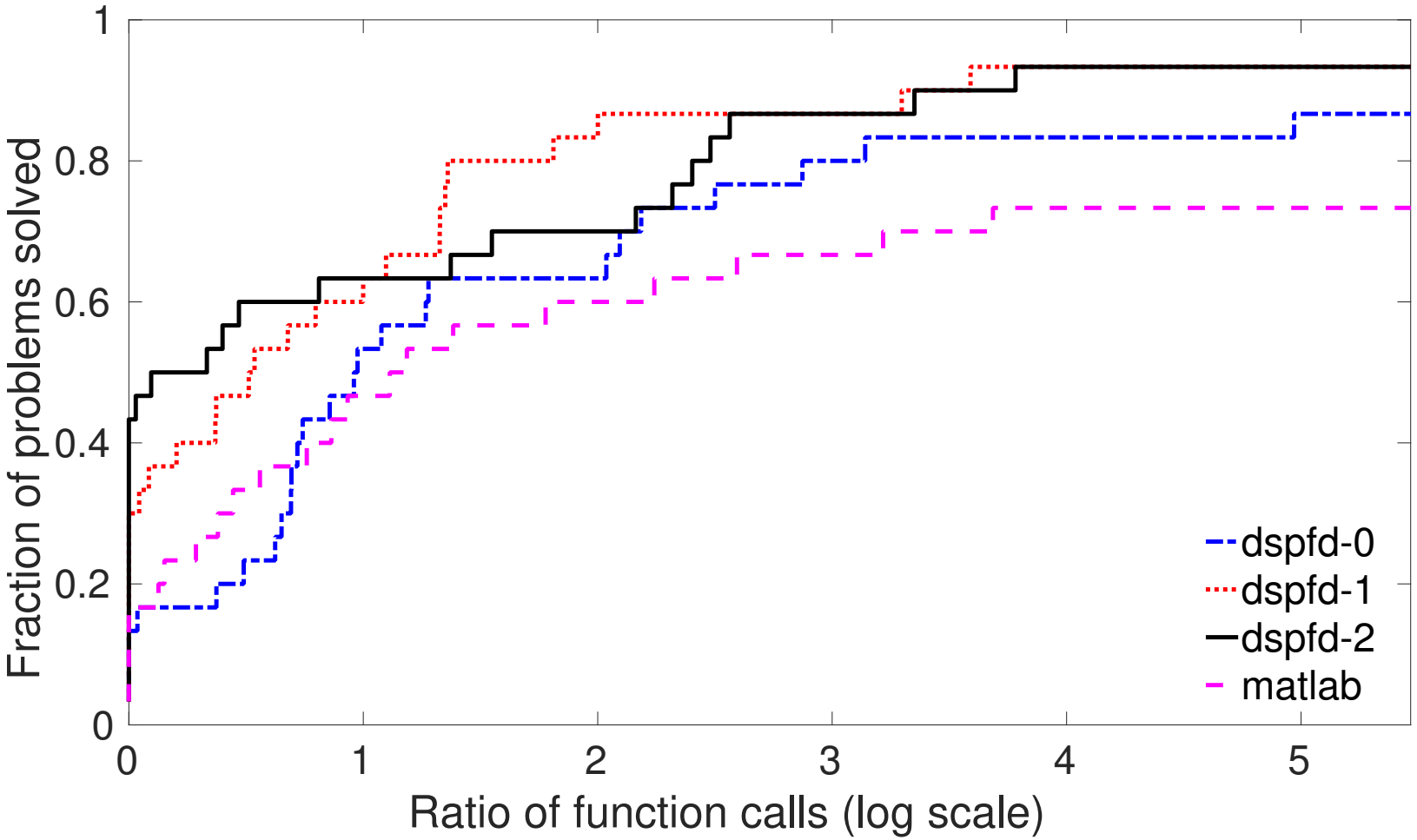
Profiles for bound-constrained problems

- Performance on 63 problems with bounds, small dimensions: $2 \leq n \leq 20$.



Profiles for bound-constrained problems (2)

- Performance on 31 problems with bounds constraints, **larger dimensions:** $20 \leq n \leq 52$.



Extension to general linear constraints

Direction generation technique

- Linear equalities: Reduce the problem to the null space of the constraint matrix.
- **Linear inequalities:** Replace $D_{\oplus} \cap T_k$ by a set of positive generators for T_k. Replace $D_{\oplus} \cap T_k^c$ by a set of positive generators for T_k^c.

Profiles for linearly constrained problems

- Performance on 106 problems with linear constraints: $2 \leq n \leq 96$.