Direct Search Based on Probabilistic Descent

Luis Nunes Vicente
University of Coimbra

January 23, 2014 — University of Oxford

http//www.mat.uc.pt/~lnv

Problem setting

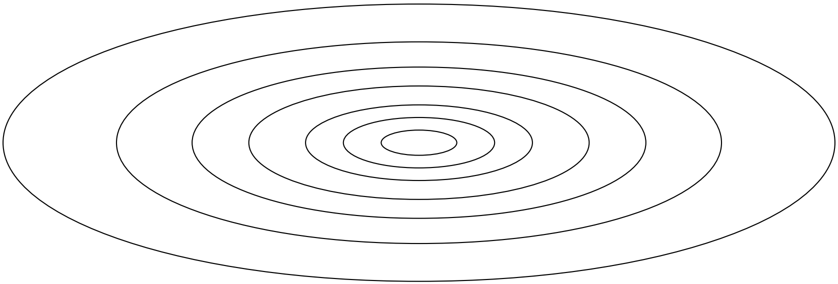
Unconstrained optimization

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

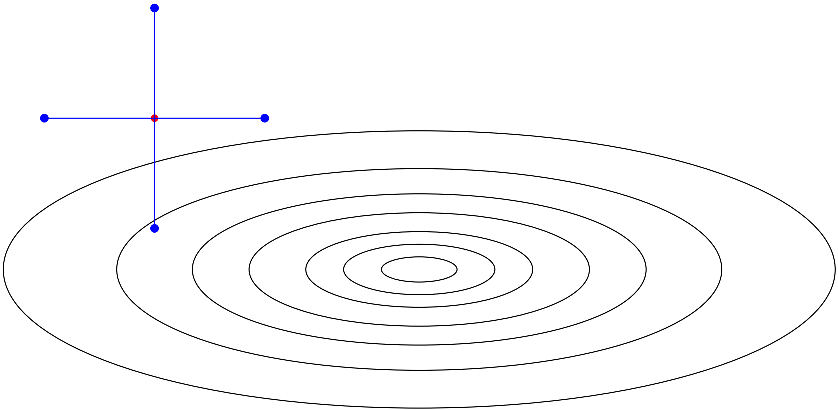
$$f : \mathbb{R}^n \to \mathbb{R}$$

f is bounded from below and differentiable
∇f is Lipschitz continuous but unavailable

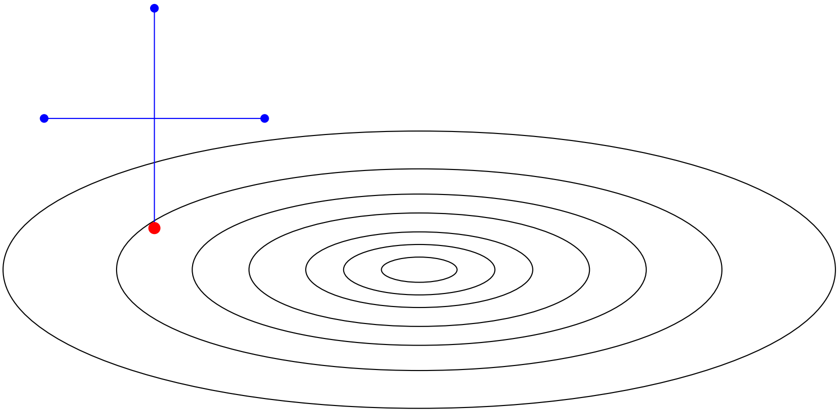
Coordinate search (poll step)



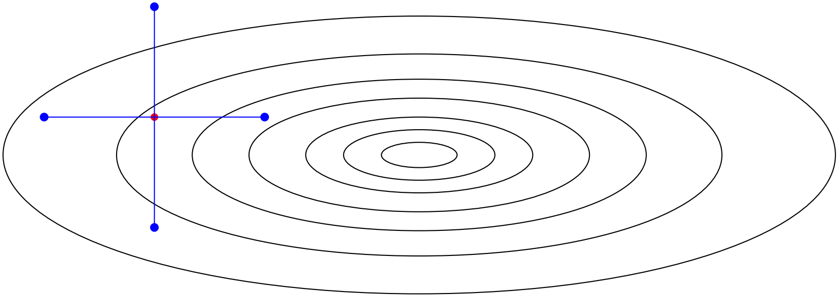
Coordinate search (poll step)



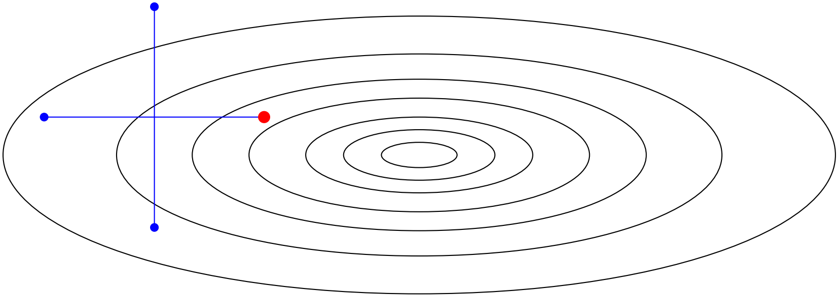
Coordinate search (poll step)



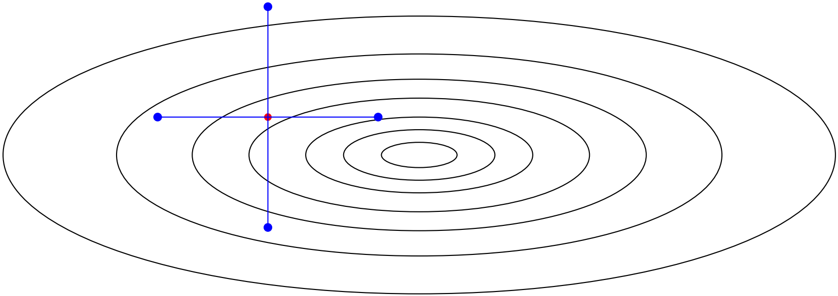
Coordinate search (poll step)



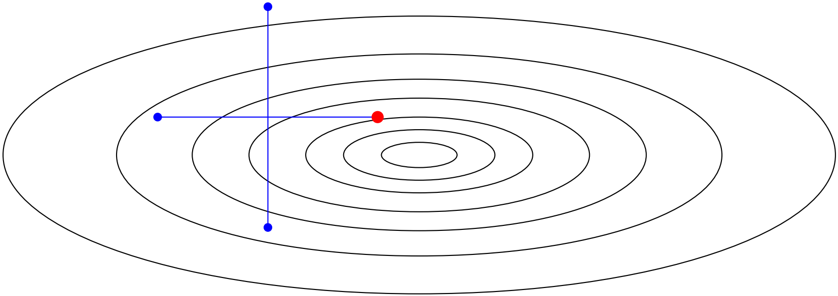
Coordinate search (poll step)



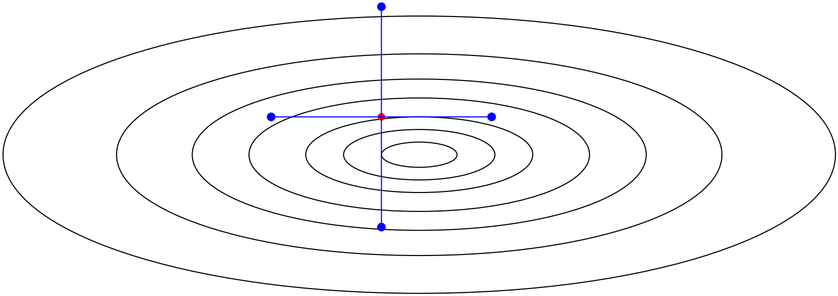
Coordinate search (poll step)



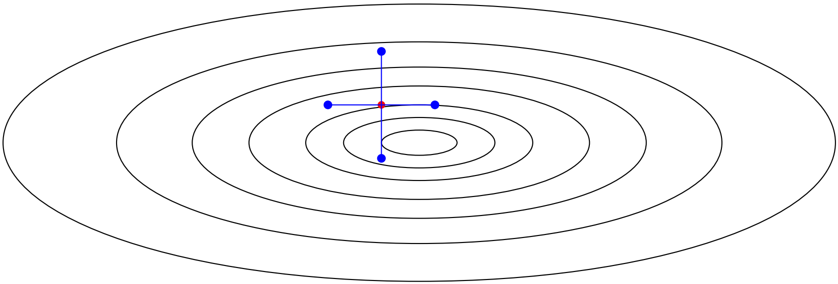
Coordinate search (poll step)

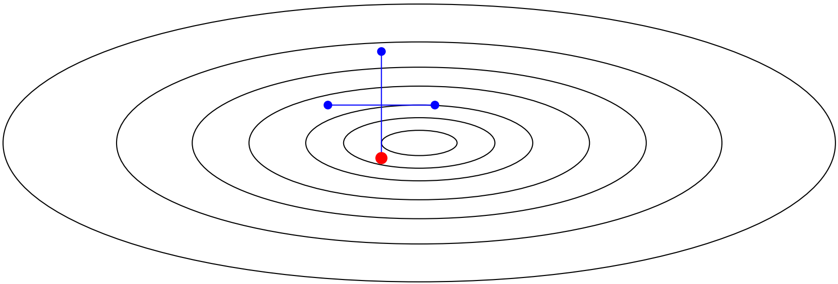


Coordinate search (poll step)

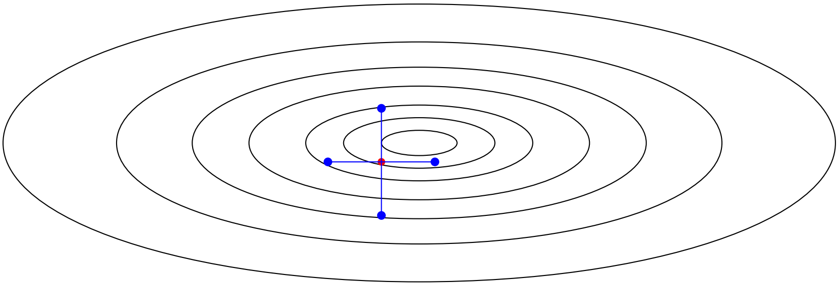


Coordinate search (poll step)

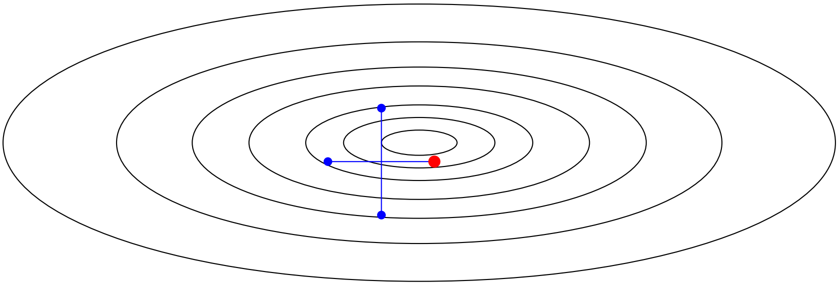




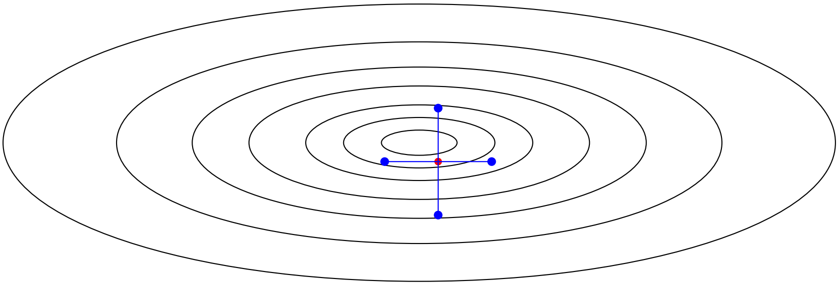
Coordinate search (poll step)



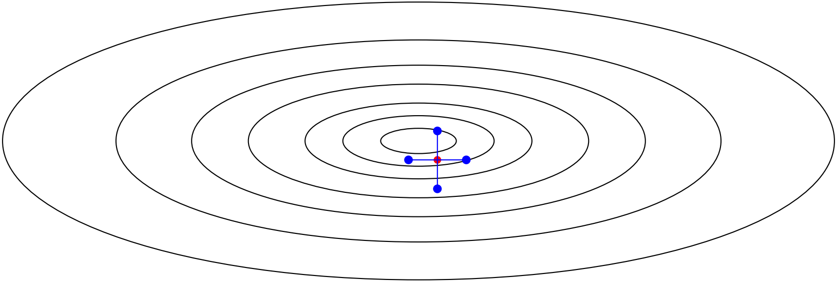
Coordinate search (poll step)



Coordinate search (poll step)



Coordinate search (poll step)



Direct search (DS)

Choose: x_0, α_0, $\gamma \in [1, \infty)$, $\theta \in (0, 1)$, and a forcing function ρ.

Direct search (DS)

Choose: x_0, α_0, $\gamma \in [1, \infty)$, $\theta \in (0, 1)$, and a forcing function ρ.

For $k = 0, 1, 2, \ldots$

- **Search step (optional)**

Direct search (DS)

Choose: x_0, α_0, $\gamma \in [1, \infty)$, $\theta \in (0, 1)$, and a forcing function ρ.

For $k = 0, 1, 2, \ldots$

- **Search step (optional)**
- **Poll step:** Select a set D_k of directions, and seek $d_k \in D_k$:

$$f(x_k + \alpha_k d_k) \; < \; f(x_k) - \rho(\alpha_k).$$

Direct search (DS)

Choose: x_0, α_0, $\gamma \in [1, \infty)$, $\theta \in (0, 1)$, and a forcing function ρ.

For $k = 0, 1, 2, \ldots$

- **Search step (optional)**
- **Poll step:** Select a set D_k of directions, and seek $d_k \in D_k$:

$$f(x_k + \alpha_k d_k) \; < \; f(x_k) - \rho(\alpha_k).$$

  If d_k is found, the iteration is successful. Otherwise, it is unsuccessful.

Direct search (DS)

Choose: x_0, α_0, $\gamma \in [1, \infty)$, $\theta \in (0, 1)$, and a forcing function ρ.

For $k = 0, 1, 2, \ldots$

- **Search step (optional)**
- **Poll step:** Select a set D_k of directions, and seek $d_k \in D_k$:

$$f(x_k + \alpha_k d_k) \ < \ f(x_k) - \rho(\alpha_k).$$

  If d_k is found, the iteration is successful. Otherwise, it is unsuccessful.

- Update the new iterate x_{k+1} (stay at x_k if unsuccessful).

Direct search (DS)

Choose: x_0, α_0, $\gamma \in [1, \infty)$, $\theta \in (0, 1)$, and a forcing function ρ.

For $k = 0, 1, 2, \ldots$

- **Search step (optional)**
- **Poll step:** Select a set D_k of directions, and seek $d_k \in D_k$:

$$f(x_k + \alpha_k d_k) \; < \; f(x_k) - \rho(\alpha_k).$$

  If d_k is found, the iteration is successful. Otherwise, it is unsuccessful.

- Update the new iterate x_{k+1} (stay at x_k if unsuccessful).
- Update the step size α_{k+1}.

  $\alpha_{k+1} = \gamma\alpha_k$ if successful, $\alpha_{k+1} = \theta\alpha_k$ if unsuccessful.

More ...

A forcing function ρ is a positive and monotonically nondecreasing function such that

$$\lim_{\alpha \downarrow 0} \frac{\rho(\alpha)}{\alpha} \;=\; 0.$$

More ...

A forcing function ρ is a positive and monotonically nondecreasing function such that

$$\lim_{\alpha \downarrow 0} \frac{\rho(\alpha)}{\alpha} = 0.$$

In this talk:

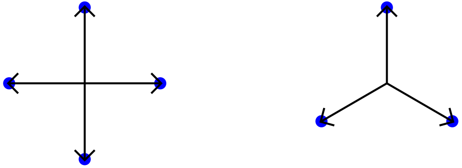
$$\rho(\alpha) = \frac{\alpha^2}{2}$$

$$\alpha_0 = 1 \qquad \text{(initial stepsize)}$$

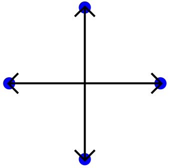
$$\gamma = 2 \qquad \text{(increasing factor)}$$

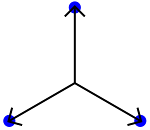
$$\theta = \frac{1}{2} \qquad \text{(decreasing factor)}$$

- Positive spanning set (PSS)



- Positive spanning set (PSS)

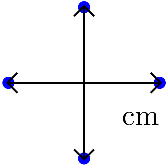




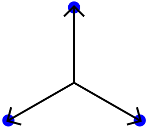
- Cosine measure of a PSS D

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

- Positive spanning set (PSS)



$\mathrm{cm} = \frac{1}{\sqrt{n}}$

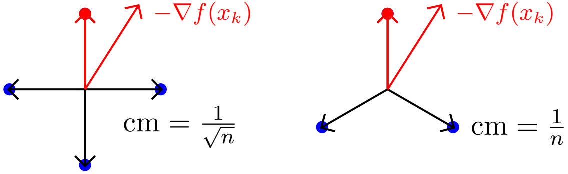


$\mathrm{cm} = \frac{1}{n}$

- Cosine measure of a PSS D

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

- Positive spanning set (PSS)



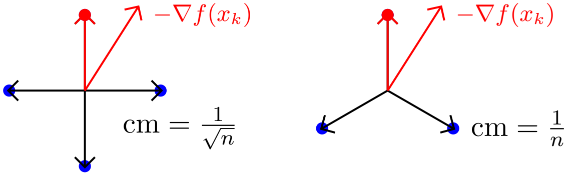
- Cosine measure of a PSS D

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

- 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 $\nabla f(x_k) \neq 0$.

  $\Longrightarrow \alpha_k$ small leads to success!

Deterministic approach (analysis)

If $\{D_k\}$ is a sequence of positive spanning sets with cosine measures bounded away from zero:

Deterministic approach (analysis)

If $\{D_k\}$ is a sequence of positive spanning sets with cosine measures bounded away from zero:

Global convergence (Torczon 1997, Kolda, Lewis, and Torczon 2003)

- $\liminf_{k\to\infty} \|\nabla f(x_k)\| = 0$.
- $\lim_{k\to\infty} \|\nabla f(x_k)\| = 0$ *if complete polling is performed.*

Deterministic approach (analysis)

If $\{D_k\}$ is a sequence of positive spanning sets with cosine measures bounded away from zero:

Global convergence (Torczon 1997, Kolda, Lewis, and Torczon 2003)

- $\liminf_{k\to\infty} \|\nabla f(x_k)\| = 0$.
- $\lim_{k\to\infty} \|\nabla f(x_k)\| = 0$ *if complete polling is performed.*

Global rate and worst case complexity (Vicente 2013)

- $\min_{0\leq\ell\leq k} \|\nabla f(x_\ell)\| = \mathcal{O}(1/\sqrt{k})$.

Deterministic approach (analysis)

If $\{D_k\}$ is a sequence of positive spanning sets with cosine measures bounded away from zero:

Global convergence (Torczon 1997, Kolda, Lewis, and Torczon 2003)

- $\liminf_{k\to\infty} \|\nabla f(x_k)\| = 0$.
- $\lim_{k\to\infty} \|\nabla f(x_k)\| = 0$ *if complete polling is performed.*

Global rate and worst case complexity (Vicente 2013)

- $\min_{0\leq\ell\leq k} \|\nabla f(x_\ell)\| = \mathcal{O}(1/\sqrt{k})$.
- $\|\nabla f(x_k)\|$ *is driven under* ϵ *within* $\mathcal{O}(\epsilon^{-2})$ *iterations.*

A more insightful look ...

If derivatives were available, it would have been sufficient to require

Descent condition

$$\mathrm{cm}\left(D_k, -\nabla f(x_k)\right) \;\geq\; \kappa \;>\; 0$$

A more insightful look ...

If derivatives were available, it would have been sufficient to require

Descent condition

$$\mathrm{cm}\left(D_k, -\nabla f(x_k)\right) \;\geq\; \kappa \;>\; 0$$

with $\mathrm{cm}(D, v)$ being the cosine measure of D given v, defined by

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

Descent condition and successful iterations

Assume the polling directions are normalized.

Lemma

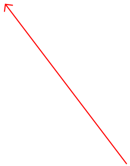
If

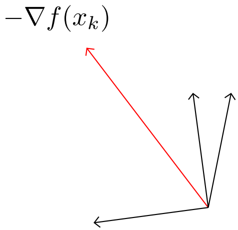
$$\mathrm{cm}\left(D_k, -\nabla f(x_k)\right) \geq \kappa \quad \textit{and} \quad \alpha_k < \frac{2\kappa\|\nabla f(x_k)\|}{\nu + 1},$$

the k*-th iteration is successful.*

The number ν is a Lipschitz constant of ∇f in $\mathbb{R}^n$.

$-\nabla f(x_k)$

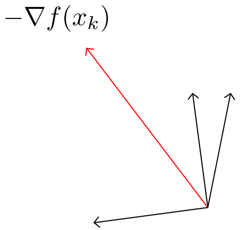




$n + 1$ random polling directions

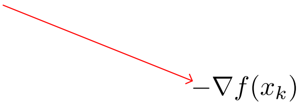
in this case not a PSS

Randomly generating 'positive spanning sets' ...

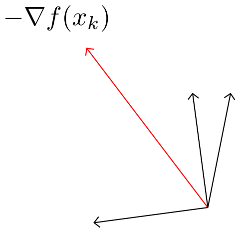


$n + 1$ random polling directions

in this case not a PSS

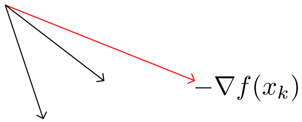


Randomly generating 'positive spanning sets' ...



$n+1$ random polling directions

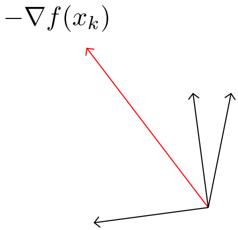
in this case not a PSS



$\leq n$ random polling directions

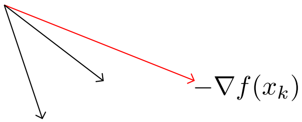
certainly not a PSS ...

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 now on, we suppose that the polling directions are not defined deterministically but generated randomly.

Probabilistic descent

From now on, we suppose that the polling directions are not defined deterministically but generated randomly.

	Iterate	Direction set
Random variables	X_k	$\mathfrak{D}_k$
Realizations	x_k	D_k

Probabilistic descent

From now on, we suppose that the polling directions are not defined deterministically but generated randomly.

	Iterate	Direction set
Random variables	X_k	$\mathfrak{D}_k$
Realizations	x_k	D_k

Definition

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

$$\mathbb{P}\Big(\mathrm{cm}(\mathfrak{D}_k, -\nabla f(X_k)) \;\geq\; \kappa \,|\, \mathfrak{D}_0, \dots, \mathfrak{D}_{k-1}\Big) \geq \; p.$$

Global convergence: The idea (by contradiction)

Intuition:

If global convergence does not hold, then $\{\mathrm{cm}(\mathfrak{D}_k, -\nabla f(X_k)) \geq \kappa\}$ probably 'rarely happens'.

Intuition:

If global convergence does not hold, then $\left\{\mathrm{cm}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) \geq \kappa\right\}$ probably 'rarely happens'.

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

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

Global convergence: The idea (by contradiction)

Without any assumption on the probabilistic behavior of $\{\mathfrak{D}_k\}$:

Lemma

$$\left\{ \liminf_{k\to\infty} \|\nabla f(X_k)\| > 0 \right\} \subset \left\{ \sum_{k=0}^{\infty} (Z_k - p_0) = -\infty \right\}.$$

Global convergence: The idea (by contradiction)

Without any assumption on the probabilistic behavior of $\{\mathfrak{D}_k\}$:

Lemma

$$\left\{ \liminf_{k \to \infty} \|\nabla f(X_k)\| > 0 \right\} \subset \left\{ \sum_{k=0}^{\infty} (Z_k - p_0) = -\infty \right\}.$$

What is this telling us?

Global convergence: The idea (by contradiction)

Without any assumption on the probabilistic behavior of $\{\mathfrak{D}_k\}$:

Lemma

$$\left\{ \liminf_{k\to\infty} \|\nabla f(X_k)\| > 0 \right\} \subset \left\{ \sum_{k=0}^{\infty} (Z_k - p_0) = -\infty \right\}.$$

$$k \left(\frac{\sum_{\ell=0}^{k-1} Z_\ell}{k} - p_0 \right) \longrightarrow -\infty,$$

and so the 'frequency' of descent would be 'eventually' below p_0.

In fact, if $\{\mathfrak{D}_k\}$ is p_0-probabilistically κ-descent, then $\left\{ \sum_{\ell=0}^{k-1} (Z_\ell - p_0) \right\}$ is a submartingale.

A submartingale $\{G_k\}$ is a sequence of random variables that are integrable ($\mathbb{E}(|G_k|) < \infty$) and that satisfy $\mathbb{E}(G_k \mid G_0, \ldots, G_{k-1}) \geq G_{k-1}$.

In fact, if $\{\mathfrak{D}_k\}$ is p_0-probabilistically κ-descent, then $\left\{ \sum_{\ell=0}^{k-1} (Z_\ell - p_0) \right\}$ is a submartingale.

A submartingale $\{G_k\}$ is a sequence of random variables that are integrable ($\mathbb{E}(|G_k|) < \infty$) and that satisfy $\mathbb{E}(G_k \mid G_0, \ldots, G_{k-1}) \geq G_{k-1}$.

Theorem

If $\{\mathfrak{D}_k\}$ is p_0-probabilistically κ-descent, then

$$\mathbb{P}\left(\liminf_{k \to \infty} \|\nabla f(X_k)\| = 0 \right) \ = \ 1.$$

Global convergence

In fact, if $\{\mathfrak{D}_k\}$ is p_0-probabilistically κ-descent, then $\left\{ \sum_{\ell=0}^{k-1} (Z_\ell - p_0) \right\}$ is a submartingale.

A submartingale $\{G_k\}$ is a sequence of random variables that are integrable ($\mathbb{E}(|G_k|) < \infty$) and that satisfy $\mathbb{E}(G_k \mid G_0, \ldots, G_{k-1}) \geq G_{k-1}$.

Theorem

If $\{\mathfrak{D}_k\}$ is p_0-probabilistically κ-descent, then

$$\mathbb{P}\left(\liminf_{k\to\infty} \|\nabla f(X_k)\| = 0 \right) \; = \; 1.$$

This analysis is a reorganization of the argument for trust regions:

- A. S. Bandeira, K. Scheinberg, and L. N. Vicente, Convergence of trust-region methods based on probabilistic models, submitted.

WCC bounds for DFO

- Non-smooth, non-convex case: $\mathcal{O}(\epsilon^{-3})$
- Non-convex case: $\mathcal{O}(\epsilon^{-2})$
- Convex case: $\mathcal{O}(\epsilon^{-1})$, global rate $1/k$

WCC bounds for DFO

- Non-smooth, non-convex case: $\mathcal{O}(\epsilon^{-3})$
- Non-convex case: $\mathcal{O}(\epsilon^{-2})$
- Convex case: $\mathcal{O}(\epsilon^{-1})$, global rate $1/k$

for Direct Search (papers by Dodangeh, Garmanjani, and Vicente)

Random Gaussian (Nesterov)

WCC bounds for DFO

- Non-smooth, non-convex case: $\mathcal{O}(\epsilon^{-3})$
- Non-convex case: $\mathcal{O}(\epsilon^{-2})$
- Convex case: $\mathcal{O}(\epsilon^{-1})$, global rate $1/k$

for Direct Search (papers by Dodangeh, Garmanjani, and Vicente)

Random Gaussian (Nesterov)

$\mathcal{O}(\epsilon^{-2}) \longrightarrow \mathcal{O}(\epsilon^{-3/2})$ using Cubic Overestimation (Cartis, Gould, Toint)

WCC bounds for DFO

- Non-smooth, non-convex case: $\mathcal{O}(\epsilon^{-3})$
- Non-convex case: $\mathcal{O}(\epsilon^{-2})$
- Convex case: $\mathcal{O}(\epsilon^{-1})$, global rate $1/k$

for Direct Search (papers by Dodangeh, Garmanjani, and Vicente)

Random Gaussian (Nesterov)

$\mathcal{O}(\epsilon^{-2}) \longrightarrow \mathcal{O}(\epsilon^{-3/2})$ using Cubic Overestimation (Cartis, Gould, Toint)

- In this talk we cover:

  S. Gratton, C. W. Royer, L. N. Vicente, Z. Zhang, Direct Search Based on Probabilistic Descent, to be submitted.

Global rate: What is desirable?

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: What is desirable?

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: What is desirable?

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: What is desirable?

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 $\|\nabla f(x_k)\| \le \epsilon$.

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

We are interested in the probabilities

Global rate

$$\mathbb{P}\big(\|\tilde{G}_k\| \le \mathcal{O}(1/\sqrt{k})\big)$$

and

Global rate: What is desirable?

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.

We are interested in the probabilities

Global rate

$$\mathbb{P}\left(\|\tilde{G}_k\| \leq \mathcal{O}(1/\sqrt{k})\right)$$

and

Worst case complexity

$$\mathbb{P}\left(K_\epsilon \leq \mathcal{O}(\epsilon^{-2})\right).$$

Global rate: The idea

Let z_ℓ denote the realization of Z_ℓ $(\ell \geq 0)$.

- Intuition: If $\|\tilde{g}_k\|$ is 'big', then $\sum_{\ell=0}^{k-1} z_\ell$ is probably 'small', thus $\sum_{\ell=0}^{k-1} z_\ell$ is possibly bounded by a (nonincreasing) function of $\|\tilde{g}_k\|$.

Global rate: The idea

Let z_ℓ denote the realization of Z_ℓ $(\ell \geq 0)$.

- Intuition: If $\|\tilde{g}_k\|$ is 'big', then $\sum_{\ell=0}^{k-1} z_\ell$ is probably 'small', thus $\sum_{\ell=0}^{k-1} z_\ell$ is possibly bounded by a (nonincreasing) function of $\|\tilde{g}_k\|$.
- In fact, we prove that

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

An auxiliary result: Behavior of the stepsize

From sufficient decrease at successful iterations

$$\rho(\alpha_k) \;\leq\; f(x_k) - f(x_{k+1}),$$

An auxiliary result: Behavior of the stepsize

From sufficient decrease at successful iterations

$$\rho(\alpha_k) \ \leq \ f(x_k) - f(x_{k+1}),$$

we obtain

$$\sum_{k \text{ successful}} \rho(\alpha_k) \ \leq \ f(x_0) - f_{\text{low}}.$$

An auxiliary result: Behavior of the stepsize

From sufficient decrease at successful iterations

$$\rho(\alpha_k) \;\leq\; f(x_k) - f(x_{k+1}),$$

we obtain

$$\sum_{k \text{ successful}} \rho(\alpha_k) \;\leq\; f(x_0) - f_{\text{low}}.$$

Lemma

For each realization of DS,

$$\sum_{k=0}^{\infty} \rho(\alpha_k) \;=\; \sum_{k=0}^{\infty} \alpha_k^2/2 \leq \frac{2}{3} + \frac{16}{3}\left[f(x_0) - f_{\text{low}}\right] \stackrel{\text{def}}{=} \beta.$$

Again, $\rho(\alpha_k) = \alpha_k^2/2$.

Number of iterations with descent

As we wanted:

Lemma

For each realization of DS,

$$\sum_{\ell=0}^{k-1} z_\ell \;\leq\; \frac{(\nu+1)^2\beta}{2\kappa^2\|\tilde{g}_k\|^2} + p_0 k.$$

Number of iterations with descent

As we wanted:

Lemma

For each realization of DS,

$$\sum_{\ell=0}^{k-1} z_\ell \;\leq\; \frac{(\nu+1)^2\beta}{2\kappa^2\|\tilde{g}_k\|^2} + p_0 k.$$

Thus,

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

Number of iterations with descent

As we wanted:

Lemma

For each realization of DS,

$$\sum_{\ell=0}^{k-1} z_\ell \;\leq\; \frac{(\nu+1)^2\beta}{2\kappa^2\|\tilde{g}_k\|^2} + p_0 k.$$

Thus,

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

A universal result

Our observation links $\mathbb{P}\big(\|\tilde{G}_k\| \leq \epsilon\big)$ to the lower tail of $\sum_{\ell=0}^{k-1} Z_\ell$.

A universal result

Our observation links $\mathbb{P}\big(\|\tilde{G}_k\| \leq \epsilon\big)$ to the lower tail of $\sum_{\ell=0}^{k-1} Z_\ell$.

Denoting

$$\pi_k(\lambda) = \mathbb{P}\left(\sum_{\ell=0}^{k-1} Z_\ell \leq \lambda\, k\right),$$

A universal result

Our observation links $\mathbb{P}(\|\tilde{G}_k\| \leq \epsilon)$ to the lower tail of $\sum_{\ell=0}^{k-1} Z_\ell$.

Denoting

$$\pi_k(\lambda) = \mathbb{P}\left(\sum_{\ell=0}^{k-1} Z_\ell \leq \lambda\, k\right),$$

one has the following universal result:

Lemma

$$\mathbb{P}(\|\tilde{G}_k\| \leq \epsilon) \;\geq\; 1 - \pi_k\left(\frac{(\nu+1)^2\beta}{2\kappa^2 k \epsilon^2} + p_0\right).$$

A universal result

Our observation links $\mathbb{P}\big(\|\tilde{G}_k\| \leq \epsilon\big)$ to the lower tail of $\sum_{\ell=0}^{k-1} Z_\ell$.

Denoting

$$\pi_k(\lambda) = \mathbb{P}\left(\sum_{\ell=0}^{k-1} Z_\ell \leq \lambda\, k\right),$$

one has the following universal result:

Lemma

$$\mathbb{P}\big(\|\tilde{G}_k\| \leq \epsilon\big) \;\geq\; 1 - \pi_k\left(\frac{(\nu+1)^2\beta}{2\kappa^2 k\epsilon^2} + p_0\right).$$

No assumption is imposed on the probabilistic behavior of $\{\mathfrak{D}_k\}$.

Chernoff bound

If $\{\mathfrak{D}_k\}$ is probabilistic descent, then π_k obeys a Chernoff type bound.

Lemma

Suppose that $\{\mathfrak{D}_k\}$ is p-probabilistically κ-descent and $\lambda \in (0, p)$. Then

$$\pi_k(\lambda) \;\leq\; \exp\left[-\frac{(p-\lambda)^2}{2p}k\right].$$

Global rate

Now we plug the Chernoff type bound into the universal result.

Theorem

Suppose that $\{\mathfrak{D}_k\}$ is p-probabilistically κ-descent with $p > p_0$ and

$$k \;\geq\; \frac{(\nu+1)^2\beta}{(p-p_0)\kappa^2\epsilon^2}.$$

Global rate

Now we plug the Chernoff type bound into the universal result.

Theorem

Suppose that $\{\mathfrak{D}_k\}$ *is* p*-probabilistically* κ*-descent with* $p > p_0$ *and*

$$k \;\geq\; \frac{(\nu+1)^2\beta}{(p-p_0)\kappa^2\epsilon^2}.$$

Then

$$\mathbb{P}\big(\|\tilde{G}_k\| \leq \epsilon\big) \;\geq\; 1 - \exp\left[-\frac{(p-p_0)^2}{8p}k\right].$$

Global rate

Theorem

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

$$\mathbb{P}\left(\|\tilde{G}_k\| \leq \left(\frac{(\nu+1)\beta^{\frac{1}{2}}}{(p-p_0)^{\frac{1}{2}}\kappa} \right) \frac{1}{\sqrt{k}} \right) \geq 1 - \exp\left[-\frac{(p-p_0)^2}{8p} k \right].$$

Global rate

Theorem

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

$$\mathbb{P}\left(\|\tilde{G}_k\| \leq \left(\frac{(\nu+1)\beta^{\frac{1}{2}}}{(p-p_0)^{\frac{1}{2}}\kappa}\right)\frac{1}{\sqrt{k}}\right) \geq 1-\exp\left[-\frac{(p-p_0)^2}{8p}k\right].$$

$\longrightarrow$ $\mathcal{O}(1/\sqrt{k})$ decaying sublinear rate for gradient holds with overwhelmingly high probability, matching the deterministic case.

Worst case complexity

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

Theorem

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

$$\mathbb{P}\left(K_\epsilon \leq \left\lceil \frac{(\nu+1)^2\beta}{(p-p_0)\kappa^2}\epsilon^{-2} \right\rceil\right) \geq 1 - \exp\left[-\frac{\beta(p-p_0)(\nu+1)^2}{8p\kappa^2}\epsilon^{-2}\right].$$

Worst case complexity

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

Theorem

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

$$\mathbb{P}\left(K_\epsilon \leq \left\lceil \frac{(\nu+1)^2\beta}{(p-p_0)\kappa^2}\epsilon^{-2} \right\rceil \right) \;\geq\; 1 - \exp\left[-\frac{\beta(p-p_0)(\nu+1)^2}{8p\kappa^2}\epsilon^{-2}\right].$$

$\longrightarrow$ $\mathcal{O}(\epsilon^{-2})$ complexity bound for # of iterations holds with overwhelmingly high probability, matching the deterministic case.

High probability iteration complexity

Proposition

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

$$k \;\geq\; \frac{3(\nu+1)^2\beta}{4(p-p_0)\kappa^2}\epsilon^{-2} - \frac{3p\ln(1-P)}{(p-p_0)^2}$$

guarantees

$$\mathbb{P}\big(\|\tilde{G}_k\| \leq \epsilon\big) \geq P.$$

Expected minimum gradient

Proposition

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

$$\mathbb{E}\left(\|\tilde{G}_k\|\right) \leq \left(\frac{(\nu+1)\beta^{\frac{1}{2}}}{(p-p_0)^{\frac{1}{2}}\kappa}\right)\frac{1}{\sqrt{k}} + \|\nabla f(x_0)\| \exp\left[-\frac{(p-p_0)^2}{8p}k\right].$$

$\longrightarrow$ $\mathcal{O}(1/\sqrt{k})$ decaying sublinear rate for $\mathbb{E}(\|\tilde{G}_k\|)$, matching the deterministic case.

Global rate implies global convergence

Proposition

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

$$\mathbb{P}\left(\inf_{k \geq 0} \|G_k\| = 0\right) \;=\; 1.$$

If the iterates never arrive at a stationary point in finite iterations, then

$$\left\{\inf_{k \geq 0} \|G_k\| = 0\right\} \;=\; \left\{\liminf_{k \to \infty} \|G_k\| = 0\right\}.$$

Without conditioning to the past ...

Proposition

If

$$\mathbb{P}\big(\mathrm{cm}(\mathfrak{D}_k, -G_k) \geq \kappa\big) \;\geq\; p$$

for each $k \geq 0$*, then*

$$\mathbb{P}\big(\|\tilde{G}_k\| \leq \epsilon\big) \;\geq\; \frac{p - p_0}{1 - p_0} - \frac{(\nu+1)^2\beta}{2(1-p_0)\kappa^2} k^{-1}\epsilon^{-2}.$$

The bound does not tend to 1 when k tends to infinity.

Practical probabilistic descent sets

For each $k \geq 0$,

- $\mathfrak{D}_k$ is independent of the previous iterations,

Practical probabilistic descent sets

For each $k \geq 0$,

- $\mathfrak{D}_k$ is independent of the previous iterations,
- $\mathfrak{D}_k$ is a set $\{\mathfrak{d}_1, \ldots, \mathfrak{d}_m\}$ of independent random vectors.
    - $\mathfrak{d}_i$ is uniformly distributed on the unit sphere,

Practical probabilistic descent sets

For each $k \geq 0$,

- $\mathfrak{D}_k$ is independent of the previous iterations,
- $\mathfrak{D}_k$ is a set $\{\mathfrak{d}_1, \ldots, \mathfrak{d}_m\}$ of independent random vectors.
  - $\mathfrak{d}_i$ is uniformly distributed on the unit sphere,
  - $\mathfrak{d}_i$ can be obtained by normalizing a vector from standard normal distribution.

$\{\mathfrak{D}_k\}$ generated in this way is probabilistically descent.

Proposition

Given $\tau \in [0, \sqrt{n}]$, $\{\mathfrak{D}_k\}$ *is* p*-probabilistically* $(\tau/\sqrt{n})$*-descent with*

$$p = 1 - \left(\frac{1}{2} + \frac{\tau}{\sqrt{2\pi}}\right)^m.$$

Practical probabilistic descent sets

$\{\mathfrak{D}_k\}$ generated in this way is probabilistically descent.

Proposition

Given $\tau \in [0, \sqrt{n}]$, $\{\mathfrak{D}_k\}$ *is* p*-probabilistically* $(\tau/\sqrt{n})$*-descent with*

$$p = 1 - \left(\frac{1}{2} + \frac{\tau}{\sqrt{2\pi}}\right)^m.$$

For instance,

$$\left.\begin{array}{rcl} m & = & 2 \\ \tau & = & \frac{1}{2} \end{array}\right\} \implies p > \frac{1}{2}$$

Worst case complexity: Dependence on the dimension

More generally, if $\{\mathfrak{D}_k\}$ is generated in this way and

$$m \; > \; \log_2 \left[1 - (\ln \theta)/(\ln \gamma)\right] \; = \; 1$$

Worst case complexity: Dependence on the dimension

More generally, if $\{\mathfrak{D}_k\}$ is generated in this way and

$$m \;>\; \log_2\left[1-(\ln\theta)/(\ln\gamma)\right] \;=\; 1$$

then it is p-probabilistically $(\tau/\sqrt{n})$-descent for some constants $p > p_0 = 1/2$ and $\tau > 0$ that are independent of n.

Worst case complexity: Dependence on the dimension

More generally, if $\{\mathfrak{D}_k\}$ is generated in this way and

$$m \;>\; \log_2\left[1-(\ln\theta)/(\ln\gamma)\right] \;=\; 1$$

then it is p-probabilistically $(\tau/\sqrt{n})$-descent for some constants $p > p_0 = 1/2$ and $\tau > 0$ that are independent of n.

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

WCC (number of iterations)

$$\mathbb{P}\left(K_\epsilon \leq \left\lceil \frac{(\nu+1)^2\beta}{(p-p_0)\tau^2}(n\epsilon^{-2})\right\rceil\right) \;\geq\; 1-\exp\left[-\frac{\beta(p-p_0)(\nu+1)^2}{8p\kappa^2}\epsilon^{-2}\right],$$

Worst case complexity: Dependence on the dimension

More generally, if $\{\mathfrak{D}_k\}$ is generated in this way and

$$m > \log_2 \left[1 - (\ln\theta)/(\ln\gamma)\right] = 1$$

then it is p-probabilistically $(\tau/\sqrt{n})$-descent for some constants $p > p_0 = 1/2$ and $\tau > 0$ that are independent of n.

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

WCC (number of iterations)

$$\mathbb{P}\left(K_\epsilon \leq \left\lceil \frac{(\nu+1)^2\beta}{(p-p_0)\tau^2}(n\epsilon^{-2}) \right\rceil\right) \geq 1 - \exp\left[-\frac{\beta(p-p_0)(\nu+1)^2}{8p\kappa^2}\epsilon^{-2}\right],$$

WCC (number of function evaluations)

$$\mathbb{P}\left(K_\epsilon^f \leq \left\lceil \frac{(\nu+1)^2\beta}{(p-p_0)\tau^2}(n\epsilon^{-2}) \right\rceil m\right) \geq 1 - \exp\left[-\frac{\beta(p-p_0)(\nu+1)^2}{8p\kappa^2}\epsilon^{-2}\right].$$

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 $[Q \;\; -Q]$.

Final remarks: General forcing function

The analysis can be extended to all forcing functions ρ satisfying the following assumption.

Assumption

There exist constants $\bar{\theta}$ and $\bar{\gamma}$ that $0 < \bar{\theta} < 1 \leq \bar{\gamma}$ such that

$$\rho(\theta\alpha) \leq \bar{\theta}\rho(\alpha), \quad \rho(\gamma\alpha) \leq \bar{\gamma}\rho(\alpha), \quad \forall\alpha > 0.$$

- Using an auxiliary function $\varphi(t) = \inf\left\{\alpha : \alpha > 0,\ \frac{\rho(\alpha)}{\alpha} + \frac{1}{2}\nu\alpha\ \geq\ t\right\}$.
- Worst case complexity in general case: $\mathcal{O}\big(1/\rho[\varphi(\kappa\epsilon)]\big)$ with overwhelmingly high probability.

Final remarks: A new proof technique

A new proof technique for establishing global rates and worst case complexity bounds for randomized algorithms for which

Final remarks: A new proof technique

A new proof technique for establishing global rates and worst case complexity 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.

Final remarks: A new proof technique

A new proof technique for establishing global rates and worst case complexity 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.

Final remarks: Trust-region case

Trust-region methods based on probabilistic models:

- Global convergence: Bandeira, Scheinberg, and Vicente 2013.

Final remarks: Trust-region case

Trust-region methods based on probabilistic models:

- Global convergence: Bandeira, Scheinberg, and Vicente 2013.
- What about a global rate?

Final remarks: Trust-region case

Trust-region methods based on probabilistic models:

- Global convergence: Bandeira, Scheinberg, and Vicente 2013.
- What about a global rate?

  One can use the same proof technique:

  - the new iterate depends on the models,
  - the models are probabilistically fully linear.

Final remarks: Trust-region case

Trust-region methods based on probabilistic models:

- Global convergence: Bandeira, Scheinberg, and Vicente 2013.
- What about a global rate?

  One can use the same proof technique:

  - the new iterate depends on the models,
  - the models are probabilistically fully linear.

  It is thus possible to obtain a global decaying rate for the gradient:

  - $\mathcal{O}(1/\sqrt{k})$, with overwhelmingly high probability.

Final remarks: Better complexity

Worst case complexity in terms of number of function evaluations:

- DS based on PSS: $\mathcal{O}(n^2\epsilon^{-2})$ (Vicente 2013).

Final remarks: Better complexity

Worst case complexity in terms of number of function evaluations:

- DS based on PSS: $\mathcal{O}(n^2\epsilon^{-2})$ (Vicente 2013).
- DS based on probabilistic descent: $\mathcal{O}(mn\epsilon^{-2})$, with overwhelmingly high probability.

Final remarks: Better complexity

Worst case complexity in terms of number of function evaluations:

- DS based on PSS: $\mathcal{O}(n^2\epsilon^{-2})$ (Vicente 2013).
- DS based on probabilistic descent: $\mathcal{O}(mn\epsilon^{-2})$, with overwhelmingly high probability.
- The second one is strictly better if m is 'smaller than' n.