Lower bounds for hit-and-run direct search

Jens Jägersküpper
Universität Dortmund, Germany

SAGA 2007 @ ETH Zürich, CH
14.9.2007

“Hit-and-run is fast and fun” (Lovász/Vempala)

What is hit-and-run all about

- Task: generate a point uniformly distributed in a convex set $K \subset \mathbb{R}^n$, where $0 < \mathrm{vol}(K) < \infty$
- Scenario: K is given by a ***K*-membership oracle**
- Approach: random walk in K such that the stationary distribution is the uniform distribution over K
- Necessary: rapidly mixing Markov chain (state space $\subseteq K$)
- Tricky: technical details

What for? Approximation of the volume of K, because:

For each deterministic algorithm there is a convex set such that after any polynomially bounded number of K-membership queries the relative approximation error is $n^{\Omega(n)}$.

In particular: Exact computation #P-hard (in this scenario).

Dyer/Frieze/Kannan (STOC '89, J.ACM '91):
An FPRAS (namely an MCMC algorithm) for the approximation of the volume with a polynomial number of queries.

Markov chains used:
originally: lattice walk (proper choice of granularity necessary)
improvement: ball walk (proper choice of radius necessary)
finally: **hit-and-run** walk

Hit-and-run in a convex set K

For a given initialization of $\boldsymbol{x} \in K \subset \mathbb{R}^n$, where $0 < \mathrm{vol}(K) < \infty$:

1. Randomly choose a line L through $\boldsymbol{x}$ (uniformly over all directions)
2. Choose $\boldsymbol{x}$ uniformly at random from the line segment $L \cap K$ (using the K-oracle). GOTO 1.

Hit-and-run mixes (at least) as rapidly as the ball/lattice walk

- without the need of finding a proper granularity/radius
- (nearly) independently of the walk's starting point in K
  Lovász/Vempala (STOC '04, SIAM J.Comp. '06)
  (initial distance from K's hull > 0)

Hit-and-run direct-search heuristic

Great! So what about direct search/optimization?

Direct search heuristic for black-box optimization of $f\colon \mathbb{R}^n \to \mathbb{R}$:

Instead of choosing a random point from $L \cap K$ using the K-oracle, do a **line search** along L using the oracle for f-evals.

For a given initialization of $\boldsymbol{x} \in \mathbb{R}^n$:

① Randomly choose a line L through $\boldsymbol{x}$, namely uniformly over all directions

② Compute an $\boldsymbol{x}' \in L$ via some line search (using the f-oracle)

③ $\boldsymbol{x} := \boldsymbol{x}'$ and GOTO 1 (or output $\boldsymbol{x}$ if stopping is requested)

Hit-and-run direct-search heuristic

For a given initialization of $\boldsymbol{x} \in \mathbb{R}^n$:

① Randomly choose a line L through $\boldsymbol{x}$, namely uniformly over all directions

② Compute an $\boldsymbol{x}' \in L$ via some line search (using the f-oracle)

③ $\boldsymbol{x} := \boldsymbol{x}'$ and GOTO 1 (or output $\boldsymbol{x}$ if stopping is requested)

Clearly, the actual line-search strategy determines

- how good such a heuristic is
- how many f-evaluations are necessary

Goal and how to get there

The goal is a **general lower bound** for all such hit-and-run direct-search heuristics

- Assumption: f has a unique optimum point $\boldsymbol{x}^* \in \mathbb{R}^n$
- Consider the approximation error in the search space, i.e. the Euclidean distance from $\boldsymbol{x}^*$
- Obvious: at least one f-evaluation per line search
- Best case is **perfect line search:** the point in L with minimum distance from $\boldsymbol{x}^*$ is chosen as $\boldsymbol{x}'$ (hypothetical)

f-evaluations $\geq$ # line searches $\geq$ # perfect line searches

---

Background/Motivation

Modification of a very basic (and old) direct search heuristic, namely the (1+1) Evolution Strategy (Rechenberg/Schwefel 1965)

which uses so-called Gaussian mutations
adapted by the so-called 1/5-(success-)rule

for the function $\text{SPHERE}\colon \mathbb{R}^n \to \mathbb{R}$ with $\text{SPHERE}(\boldsymbol{x}) = |\boldsymbol{x}|^2$
(yes, a toy problem, the simplest quadratic).

(1+1) Evolution Strategy

(1+1) ES

```
⓪ Init $\boldsymbol{x} \in \mathbb{R}^n$ (starting point)

① Randomly generate $\boldsymbol{m} \in \mathbb{R}^n$
  (depending on the course of the optimization)
② Generate mutant $\boldsymbol{x}' := \boldsymbol{x} + \boldsymbol{m}$.
③ IF $f(\boldsymbol{x}') \leq f(\boldsymbol{x})$ THEN $\{\boldsymbol{x} := \boldsymbol{x}'\}$

④ IF stopping criterion THEN output $\boldsymbol{x}$ ELSE GOTO 1
```

(1+1) Evolution Strategy

(1+1) ES with 1/5-rule and scaled Gaussian mutations

```
⓪ Init $\boldsymbol{x} \in \mathbb{R}^n$ (starting point) and $\sigma > 0$ (mutation strength);
  $g := 0$;  $b := 0$;
① Randomly generate $\boldsymbol{m} \in \mathbb{R}^n$
  by $\boldsymbol{m} := \sigma \cdot \tilde{\boldsymbol{m}}$   where   $\tilde{\boldsymbol{m}} \sim \big(N_1(0;1), \ldots, N_n(0;1)\big)$
② Generate mutant $\boldsymbol{x}' := \boldsymbol{x} + \boldsymbol{m}$.
③ IF $f(\boldsymbol{x}') \leq f(\boldsymbol{x})$ THEN $\{\boldsymbol{x} := \boldsymbol{x}'; g{+}{+}\}$ ELSE $\{b{+}{+}\}$
④ IF $g + b = 5n$ THEN {
      IF $g > (g + b) \cdot 1/5$ THEN $\sigma \mathrel{*}= 2$ ELSE $\sigma \mathrel{/}= 2$;
      $g := 0$;  $b := 0$ }
⑤ IF stopping criterion THEN output $\boldsymbol{x}$ ELSE GOTO 1
```

Gaussian Mutations

Each of the n components of the vector $\tilde{\boldsymbol{m}}$ is independently standard-normally distributed.

Density at $\boldsymbol{x} \in \mathbb{R}^n$ equals

$$\prod_{i=1}^{n} \frac{\exp(-{x_i}^2/2)}{\sqrt{2\pi}} = \frac{\exp\left(\sum_{i=1}^{n} -{x_i}^2/2\right)}{\sqrt{2\pi}} = \frac{\exp(-|\boldsymbol{x}|^2/2)}{\sqrt{2\pi}}$$

$\Longrightarrow$ same density for vectors of same length $\Longrightarrow$ isotropy

The r.v. $|\tilde{\boldsymbol{m}}|$ is a χ-distribution with n degrees of freedom, its density at $\ell \in \mathbb{R}_{\geq 0}$ equals $\frac{\ell^{n-1}}{\exp(\ell^2/2)} \cdot \frac{2}{2^{n/2} \cdot \Gamma(n/2)}$.

$$\mathrm{E}\left[\,|\tilde{\boldsymbol{m}}|\,\right] \asymp \sqrt{n},\ \mathrm{Var}\left[\,|\tilde{\boldsymbol{m}}|\,\right] \leq 1/2,\ \mathrm{Prob}\left\{|\tilde{\boldsymbol{m}}| \neq \Theta(\sqrt{n})\right\} = O(\tfrac{1}{n})$$

Isotropic Distributions

Definition

Let a vector $\boldsymbol{x}$ be distributed according to some distribution F over $\mathbb{R}^n$.

Then F is *isotropic* (or *spherically symmetric*) if it is invariant w. r. t. orthonormal transformations, i. e., for any orthogonal matrix $\boldsymbol{M}$ (i. e., $\boldsymbol{M}^\top \boldsymbol{M} = \boldsymbol{I}$), $\boldsymbol{x}$ and $\boldsymbol{M}\boldsymbol{x}$ are equidistributed.

Consequently, the random direction of an isotropically distributed vector $\boldsymbol{x}$ and its random length are independent;

moreover, the direction of $\boldsymbol{x}$ is uniformly distributed.

Background/Motivation (2)

An isotropic mutation is an "oblivious" line search
which samples exactly one point from the halve line (which starts at $\boldsymbol{x}$ and poins in a uniformly random direction).

Modification of the (1+1) ES: If for a chosen scaled Gaussian mutation $\boldsymbol{m} \in \mathbb{R}^n$ the point $\boldsymbol{x} + \boldsymbol{m}$ improves upon $\boldsymbol{x}$, then check whether $\boldsymbol{x} + 2\boldsymbol{m}$ is even better.

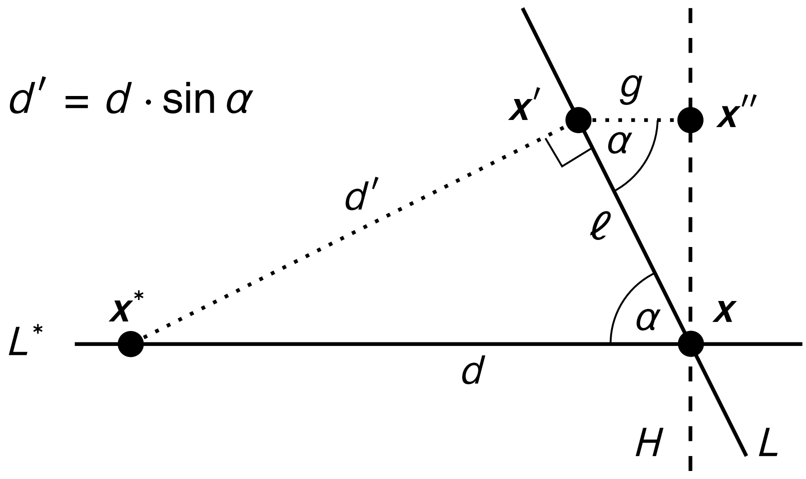
Idea: Fully exploit a good direction once it is found

Observation: Does not help at all – even worse:
more samples/f-evaluations needed

Illustration: perfect line search

For a given optimum $\boldsymbol{x}^* \in \mathbb{R}^n$ and a given init of $\boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{x}^*\}$:

1. Randomly choose a line L through $\boldsymbol{x}$ (uniformly over all directions)
2. Choose $\boldsymbol{x} \in L$ with minimum distance from $\boldsymbol{x}^*$. GOTO 1.



$d' = d \cdot \sin\alpha$

$g = \ell \cdot \cos\alpha$

$$\boxed{\frac{d'}{d} = \sqrt{1 - (g/\ell)^2}}$$

Technical details ...

Density of r.v. g/ℓ at $x \in [0;1]$ for $n \geq 4$ is $(1-x^2)^{(n-3)/2}/\Psi$, where $\Psi := \int_0^1 (1-x^2)^{(n-3)/2} \mathrm{d}x$ (normalization).

$$\begin{aligned} \mathrm{E}\left[d'/d\right] &= \int_0^1 \sqrt{1-x^2} \cdot (1-x^2)^{(n-3)/2}/\Psi \, \mathrm{d}x \\ &= \int_0^1 (1-x^2)^{n/2-1}/\Psi \, \mathrm{d}x \\ &= \frac{\Gamma(n/2) \cdot \Gamma(n/2)}{\Gamma(n/2-1/2) \cdot \Gamma(n/2+1/2)} \end{aligned}$$

Hence, for $n \geq 4$,

$$1 - \frac{1}{n} \quad < \quad \mathrm{E}\left[d'/d\right] \quad < \quad 1 - \frac{1}{2n}$$

Convergence (perfect line search)

Theorem

The (hypothetical) hit-and-run heuristic, which uses *perfect line search*, converges linearly (in the search space $\mathbb{R}^n$) at an expected rate of $1 - \beta(n)/n$, where $\beta \colon \mathbb{N} \to (0.5; 1)$.

Korollar

For **any** hit-and-run direct search heuristic: The reduction of the approximation error (in the search space) in a single step/line search is less than the nth fraction (in expectation).

Drift

Conclusion from the *expected* one-step gain
about the *expected* number of steps
necessary for a *fixed* reduction of the approximation error ?

Lemma

Let $X_1, X_2, \ldots$ denote r.v.s with bounded range and T the r.v. defined by $T := \min\{\, t \mid X_1 + \cdots + X_t \geq g\}$ for a given $g > 0$. Given that T is a stopping time (i. e., the event $\{T = t\}$ depends only on $X_1, \ldots, X_t$) and $\mathrm{E}[T] < \infty$,

if $\mathrm{E}\left[X_i \mid T \geq i\right] \leq u \neq 0$ for $i \in \mathbb{N}$, then $\mathrm{E}[T] \geq g/u$,

if $\mathrm{E}\left[X_i \mid T \geq i\right] \geq \ell > 0$ for $i \in \mathbb{N}$, then $\mathrm{E}[T] \leq \mathrm{E}[X_1 + \cdots + X_T]/\ell$.

Lower bound on the expected number of steps

Apply the previous lemma to halving the approximation error, where d denotes the approx. error (i.e. distance from $\boldsymbol{x}^*$).

$g := d/2$ and $u := d/n$, so that $\mathrm{E}[T] \geq g/u = 0.5\,n$.

By linearity of expectation:

Theorem

Hit-and-run heuristic with perfect line search in $\mathbb{R}^n$ with $n \geq 4$. Let $b \colon \mathbb{N} \to \mathbb{N}$. The expected number of steps until the distance from $\boldsymbol{x}^*$ is less than the $2^{-b(n)}$-fraction of the initial one is bounded below by $b(n) \cdot 0.5\,n - b(n) + 1$.

Stopover

So, the **expected** number of steps to halve the approximation error is at least $0.5\,n$.

There could be a good chance of getting by with considerably fewer steps.

Actually, there is only an exponentially small (in n) chance of getting by with considerably fewer steps.

Sharper bounds: Hoeffding's bound

Theorem

Let $X_1, \ldots, X_k$ denote independent random variables, each with bounded range so that $a_i \le X_i \le b_i$ with $a_i < b_i$ for $i \in \{1, \ldots, k\}$. Let $S := X_1 + \cdots + X_k$. Then for any $x > 0$ $\mathrm{Prob}\{S \ge \mathrm{E}[S] + x\} \le \exp(-2\,x^2 / \sum_{i=1}^{k}(b_i - a_i)^2)$.

$\mathrm{Prob}\{S \ge \mathrm{E}[S] + x\} \le \exp(-2 \cdot (x/b)^2/k)$ in case $X_i \in [0; b]$

Steps of a (hit-and-run) direct search not independent (in g.)!

Stochastic Dominance

Given r.v.s X and Y. Then X stochastically dominates Y, notation „$X \succ Y$", iff $\forall a \in \mathbb{R}$: $\text{Prob}\{X \le a\} \le \text{Prob}\{Y \le a\}$.

- $X \succ Y$ and $\mathrm{E}[X] < \infty \implies \mathrm{E}[X] \ge \mathrm{E}[Y]$
- $X \succ Y \succ Z \implies X \succ Z$ (transitivity)
- $X \succ X'$ and $Y \succ Y' \implies X + Y \succ X' + Y'$

Let Δ_i denote the gain in the ith step of a direct search heuristic.

If we "find" independent r.v.s X_i such that $X_i \succ \Delta_i$ then

$$\Delta_1 + \cdots + \Delta_k \prec X_1 + \cdots + X_k := S_k$$

so that we can apply Hoeffding's bound to S_k

Actual application of Hoeffding's bound

For a given optimum $\boldsymbol{x}^* \in \mathbb{R}^n$ and a given init of $\boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{x}^*\}$:

① Randomly choose a line L through $\boldsymbol{x}$
(uniformly over all directions)

② Choose $\boldsymbol{x} \in L$ with minimum distance from $\boldsymbol{x}^*$. GOTO 1.

Scale invariance:

Δ_i / d_{i-1} has the same distribution as Δ_j / d_{j-1},
where d_k denotes the distance from $\boldsymbol{x}^*$ after the kth step

In other words: Independently of the gain in the first step, the gain in the second (3rd, ...) step cannot be "better", because $d_0 \ge d_1 \ge d_2 \ge \ldots$ so that $\Delta_1 \succ \Delta_2 \succ \Delta_3 \succ \ldots$

We can choose i.i.d. copies of the r.v. Δ_1 as the r.v.s X_i
– and then we can apply Hoeffding's bound

“Sharp” lower bound on the number of steps

Theorem

Hit-and-run direct search heuristic with perfect line search in $\mathbb{R}^n$. Let $b\colon \mathbb{N} \to \mathbb{N}$ such that $b(n) = \text{poly}(n)$.

With probability $1 - e^{-\Omega(n^{1/3})}$ more than $b(n) \cdot 0.4\, n$ steps are necessary to reduce the approximation error below the $2^{-b(n)}$th fraction of the initial one.

Clearly, perfect line search is hypothetical, BUT:

Via induction on the #steps we can prove that no other line-search strategy can reduce the approx. error „faster” (in the sense of stochastic dominance).

$\Longrightarrow$ The lower bounds presented hold for **any** hit-and-run direct-search heuristic !

Conclusion

Any hit-and-run direct-search heuristic converges at best linearly at an expected rate worse (i.e. larger) than $1 - 1/n$

#steps to halve the approx. error grows linearly in the search space dimension, even with very high probability.

The simple (1+1) Evolution Strategy with Gaussian mutations adapted by the 1/5-rule gets by with $O(n)$ steps with very high probability (at least in the toy-function scenario)

Out of the question, for quadratics in general, this is bad in comparison, e.g., with BFGS or generalized/nonlinear CG

Nevertheless: Strong lower bound for a large class of direct search heuristics, in particular for *several* adaptation mechanisms for (1+1) evolutionary algorithms