Minimization by Random Search Techniques
by Solis and Wets

and

an Intro to Sampling Methods

Presenter: Michele Aghassi

October 27, 2003

This presentation is based on: Solis, F. J., and R. J-B. Wets. *Minimization by Random Search Techniques. Mathematical Operations Research* 6, 1981, pp. 19-30.

Recap of Past Sessions

- LP
  - Kalai (1992, 1997)
    - use randomized pivot rules
  - Motwani and Raghavan (1995), Clarkson (1998, 1995)
    - solve on a random subset of constraints, recursively
  - Dunagan and Vempala (2003): LP Feasibility ($\mathbf{Ax} \geq \mathbf{0}, \mathbf{0} \neq \mathbf{0}$)
    - Generate random vectors and test for feasibility
    - If not, try moving in deterministic (w.r.t. random vector already selected) direction to achieve feasibility
- NLP
  - Storn and Price (1997): Unconstrained NLP
    - Heuristic
    - Select random subsets of solution population vectors
    - Perform addition, subtraction, component swapping and test for obj func improvement

Motivation

What about provably convergent algorithms for constrained NLPs?

- Random search techniques first proposed in the 1950s
- pre-1981 proofs of convergence were highly specific and involved
- Solis and Wets, 1981: Can we give more general sufficient conditions for convergence, unifying the past results in the literature?
- Solis and Wets paper interesting more from a unifying theoretical standpoint
- Computational results of the paper relatively unimpressive

Outline

- Part I: Solis and Wets paper
  - Motivation for using random search
  - Appropriate goals of random search algorithms
  - Conceptual Algorithm encompassing several concrete examples
  - Sufficient conditions for global search convergence, and theorem
  - Local search methods and sufficient conditions for convergence, and theorem
  - Defining stopping criteria
  - Some computational results
- Part II: Intro to Sampling Methods
  - Traditional Methods
  - Hit-and-run algorithm

Why Use Random Search Techniques?

Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, $S \subseteq \mathbb{R}^n$.

$$\begin{array}{lll} \text{(P)} & \min & f(\mathbf{x}) \\ & \text{s.t.} & \mathbf{x} \in S \end{array}$$

- Function characteristics difficult to compute (e.g. gradients, etc.)
- Function is "bumpy"
- Need global minimum, but there are lots of local minima
- Limited computer memory

What is an Appropriate Goal?

- Problems
  - Global min may not exist
  - Finding min may require exhaustive examination (e.g. min occurs at point at which f singularly discontinuous)
- Response

  Definition 1. *α is the **Essential Infimum** of f on S iff*

$$\alpha \quad = \quad \inf\left\{t \mid v(\mathbf{x} \in S \mid f(\mathbf{x}) < t) > 0\right\},$$

  *where v denotes n-dimensional volume or Lebesgue measure. **Optimality region** for P is given by*

$$R_{\epsilon,M} \quad = \quad \begin{cases} \{\mathbf{x} \in S \mid f(\mathbf{x}) < \alpha + \epsilon\}, & \alpha \textit{ finite} \\ \{\mathbf{x} \in S \mid f(\mathbf{x}) < -M\}, & \alpha = -\infty, \end{cases}$$

  for a given "big" $M > 0$

What is Random Search?

Conceptual Algorithm:

1. Initialize: Find $\mathbf{x}^0 \in S$. Set $k := 0$
2. Generate $\xi^k \in \mathbb{R}^n$ (random) from distribution μ_k
3. Set $\mathbf{x}^{k+1} = D(\mathbf{x}^k, \xi^k)$. Choose μ_{k+1}. Set $k := k + 1$. Go to step 1.

$$\mu_k(A) \quad = \quad P\left(\mathbf{x}^k \in A \;\middle|\; \mathbf{x}^0, \mathbf{x}^1, \ldots, \mathbf{x}^{k-1}\right)$$

This captures both

- Local search $\implies$ $\text{supp}(\mu_k)$ is bounded and $v(S \cap \text{supp}(\mu_k)) < v(S)$
- Global search $\implies$ $\text{supp}(\mu_k)$ is such that $v(S \cap \text{supp}(\mu_k)) = v(S)$

Sufficient Conditions for Convergence

(H1) D s.t. $\left\{f(\mathbf{x}^k)\right\}_{k=0}^{\infty}$ nonincreasing

$$\begin{aligned} & & f(D(\mathbf{x},\xi)) &\leq f(\mathbf{x}) \\ \xi \in S & \implies & f(D(\mathbf{x},\xi)) &\leq \min\{f(\mathbf{x}), f(\xi)\} \end{aligned}$$

(H2) Zero probability of repeatedly missing any positive-volume subset of S.

$$\forall A \subseteq S \text{ s.t. } v(A) > 0, \qquad \prod_{k=0}^{\infty} (1 - \mu_k(A)) \quad = \quad 0$$

i.e. sampling strategy given by μ_k cannot consistently ignore a part of S with positive volume (Global search methods satisfy (H2))

Example Satisfying (H1) and (H2), I

Due to Gaviano [2].

$$\begin{aligned} D(\mathbf{x}^k, \xi^k) &= (1-\lambda_k)\mathbf{x}^k + \lambda_k \xi^k \text{ where} \\ \lambda_k &= \arg\min_{\lambda\in[0,1]} \left[f((1-\lambda)\mathbf{x}^k + \lambda\xi^k) \mid (1-\lambda)\mathbf{x}^k + \lambda\xi^k \in S \right] \end{aligned}$$

μ_k unif on n-dim sphere with center $\mathbf{x}^k$ and $r \geq 2\text{diam}(S)$.

Why?

- (H1) satisfied since $\left\{ f(\mathbf{x}^k) \right\}_{k=0}^{\infty}$ nonincreasing by construction
- (H2) satisfied because sphere contains S

Example Satisfying (H1) and (H2), II

Due to Baba *et al.* [1].

$$D(\mathbf{x}^k, \xi^k) = \begin{cases} \xi^k, & \xi^k \in S \text{ and } f(\xi^k) < f(\mathbf{x}^k) \\ \mathbf{x}^k, & \text{o.w.} \end{cases}$$

$$\mu_k \sim \mathcal{N}(\mathbf{x}^k, \mathbf{I})$$

Why?

- (H1) satisfied since $\left\{ f(\mathbf{x}^k) \right\}_{k=0}^{\infty}$ nonincreasing by construction
- (H2) satisfied because S contained in support of $\mathcal{N}(\mathbf{x}^k, \mathbf{I})$

Global Search Convergence Theorem

Theorem 1. *Suppose f measurable, $S \subseteq \mathbb{R}^n$ measurable, (H1), (H2), and $\{\mathbf{x}^k\}_{k=0}^{\infty}$ generated by the algorithm. Then*

$$\lim_{k\to\infty} P\left(\mathbf{x}^k \in R_{\epsilon,M}\right) \quad = \quad 1$$

Proof. By (H1), $\mathbf{x}^k \notin R_{\epsilon,M} \implies \mathbf{x}^\ell \notin R_{\epsilon,M}, \forall \ell < k$

$$P\left(\mathbf{x}^k \in S \backslash R_{\epsilon,M}\right) \quad \leq \quad \prod_{\ell=0}^{k-1} \left(1 - \mu_\ell(R_{\epsilon,M})\right)$$

$$P\left(\mathbf{x}^k \in R_{\epsilon,M}\right) \;=\; 1 - P\left(\mathbf{x}^k \in S \backslash R_{\epsilon,M}\right) \quad \geq \quad 1 - \prod_{\ell=0}^{k-1} \left(1 - \mu_\ell(R_{\epsilon,M})\right)$$

$$1 \geq \lim_{k\to\infty} P\left(\mathbf{x}^k \in R_{\epsilon,M}\right) \quad \geq \quad 1 - \lim_{k\to\infty} \prod_{\ell=0}^{k-1} \left(1 - \mu_\ell(R_{\epsilon,M})\right) = 1,$$

where last equality follows from (H2). □

Local Search Methods

- Easy to find examples for which the algorithm will get trapped at local minimum
- Drastic sufficient conditions ensure convergence to optimality region, but are very difficult to verify

For instance

(H3) $\forall \mathbf{x}^0 \in S$
$L_0 = \{\mathbf{x} \in S \mid f(\mathbf{x}) \leq f(\mathbf{x}^0)\}$ is compact and
$\exists \gamma > 0$ and $\eta \in (0, 1]$ (possibly depending on $\mathbf{x}^0$) s.t., $\forall k$ and $\forall \mathbf{x} \in L_0$,

$$\mu_k \left(\left[D(\mathbf{x}, \xi) \in R_{\epsilon,M} \right] \;\cup\; \left[\mathrm{dist}(D(\mathbf{x}, \xi), R_{\epsilon,M}) < \mathrm{dist}(\mathbf{x}, R_{\epsilon,M}) - \gamma \right] \right) \quad \geq \quad \eta.$$

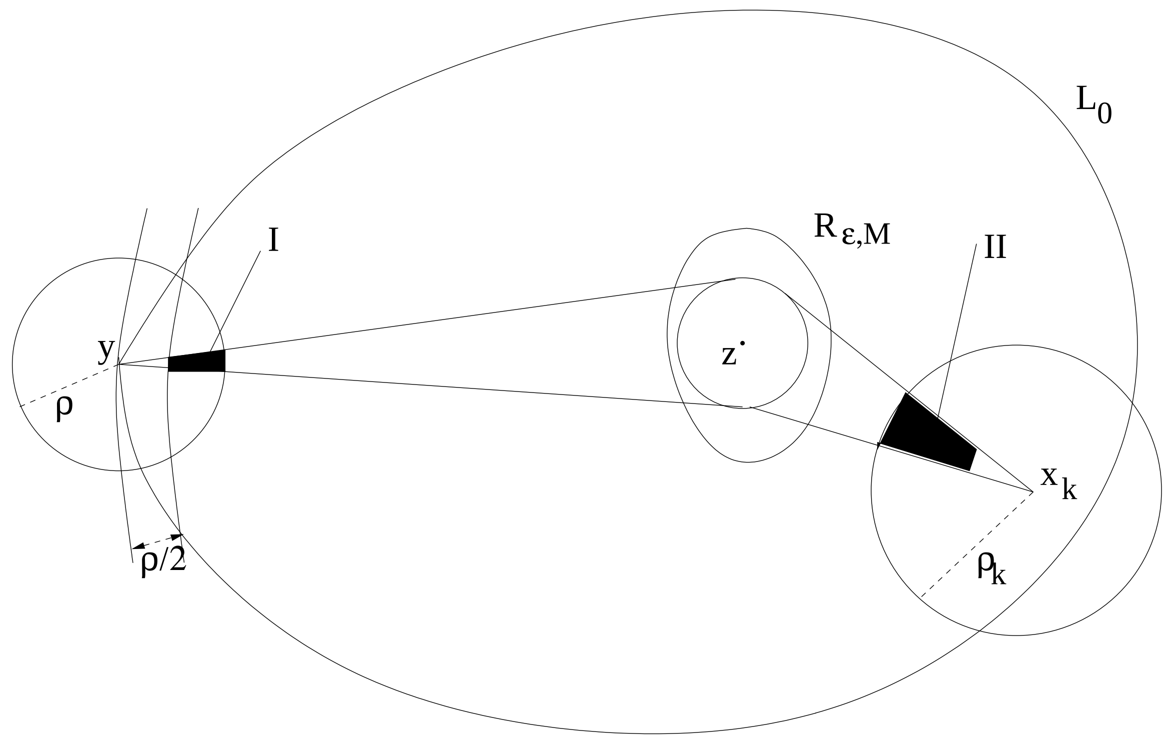
If f and S are "nice," local search methods demonstrate better convergence behavior.

Example Satisfying (H3), I

- $\text{int}(S) \neq \emptyset$
- $\forall \alpha \in \mathbb{R}$, $S \cap \{\mathbf{x} \mid f(\mathbf{x}) \leq \alpha\}$ convex and compact
  Happens whenever f quasi-convex and either S compact or f has bounded level sets
- ξ^k chosen via uniform distribution on hypersphere with center $\mathbf{x}^k$ and radius ρ_k
- ρ_k is a function of $\mathbf{x}^0, \mathbf{x}^1, \ldots, \mathbf{x}^{k-1}$ and $\xi^1, \ldots, \xi^{k-1}$ such that $\rho = \inf_k \rho_k > 0$
- $$D(\mathbf{x}^k, \xi^k) = \begin{cases} \xi^k, & \xi^k \in S \\ \mathbf{x}^k, & \text{o.w.} \end{cases}$$

Proof. L_0 compact convex since level sets are.
$R_{\epsilon,M}$ has nonempty interior since S does.
$\therefore$ can draw ball contained in interior of $R_{\epsilon,M}$.
Now take $\gamma = \frac{\rho}{2}$ and $\eta = \frac{v(\text{region I})}{v(\text{hypersphere with radius } \rho)} > 0$

Example Satisfying (H3)



$$\frac{v(\text{region II})}{v(\text{hypersphere with radius } \rho_k)} > \frac{v(\text{region I})}{v(\text{hypersphere with radius } \rho)} = \eta. \qquad \square$$

Local Search Convergence Theorem, I

Theorem 2. *Suppose f is a measurable function, $S \subseteq \mathbb{R}^n$ is a measurable, and (H1) and (H3) are satisfied. Let $\left\{\mathbf{x}^k\right\}_{k=0}^{\infty}$ be a sequence generated by the algorithm. Then,*

$$\lim_{k\to\infty} P\left(\mathbf{x}^k \in R_{\epsilon,M}\right) \quad = \quad 1.$$

Proof. Let $\mathbf{x}^0$ be the initial iterate used by the algorithm. By (H1), all future iterates in $L_0 \supseteq R_{\epsilon,M}$. L_0 is compact. Therefore $\exists p \in \mathbb{Z}$ s.t. $\gamma p > \operatorname{diam}(L_0)$.

$$\begin{aligned}
P\left(\mathbf{x}^{\ell+p} \in R_{\epsilon,M} \mid \mathbf{x}^{\ell} \notin R_{\epsilon,M}\right) \quad &= \quad \frac{P\left(\mathbf{x}^{\ell+p} \in R_{\epsilon,M},\ \mathbf{x}^{\ell} \notin R_{\epsilon,M}\right)}{P\left(\mathbf{x}^{\ell} \notin R_{\epsilon,M}\right)} \\
&\geq \quad P\left(\mathbf{x}^{\ell+p} \in R_{\epsilon,M},\ \mathbf{x}^{\ell} \notin R_{\epsilon,M}\right) \\
&\geq \quad P\Big(\mathbf{x}^{\ell} \notin R_{\epsilon,M},\ \operatorname{dist}(\mathbf{x}^k, R_{\epsilon,M}) \leq \gamma(p-(k-\ell)), \\
&\qquad\quad k = \ell, \ldots, \ell+p\Big) \\
&\geq \quad \eta^p \qquad \text{by repeated Bayes rule and (H3)}
\end{aligned}$$

Local Search Convergence Theorem, II

Claim: $P\left(\mathbf{x}^{kp} \notin R_{\epsilon,M}\right) \leq (1-\eta^p)^k, \forall k \in \{1, 2, \dots\}$

By induction

$$
\begin{array}{lrcl}
(k=1) & P\left(\mathbf{x}^p \in R_{\epsilon,M}\right) & \geq & P\left(\mathbf{x}^p \in R_{\epsilon,M},\ \mathbf{x}^0 \notin R_{\epsilon,M}\right) \ \geq \ \eta^p \\
(\text{Genl } k) & P\left(\mathbf{x}^{kp} \notin R_{\epsilon,M}\right) & = & P\left(\mathbf{x}^{kp} \notin R_{\epsilon,M} \mid \mathbf{x}^{(k-1)p} \notin R_{\epsilon,M}\right) P\left(\mathbf{x}^{(k-1)p} \notin R_{\epsilon,M}\right) \\
 & & \leq & \left[1 - P\left(\mathbf{x}^{kp} \in R_{\epsilon,M} \mid \mathbf{x}^{(k-1)p} \notin R_{\epsilon,M}\right)\right] \left(1-\eta^p\right)^{k-1} \\
 & & \leq & \left(1-\eta^p\right)\left(1-\eta^p\right)^{k-1}
\end{array}
$$

$$
\therefore\ P\left(\mathbf{x}^{kp+\ell} \in R_{\epsilon,M}\right) \ \geq \ P\left(\mathbf{x}^{kp} \in R_{\epsilon,M}\right) \ \geq \ 1 - \left(1-\eta^p\right)^k, \qquad \ell = 0, 1, \dots, p-1
$$

$\square$

Stopping Criteria

- So far, we gave a conceptual method for generating $\{\mathbf{x}^k\}_{k=0}^{\infty}$ such that $f(\mathbf{x}^k) \rightarrow$ essential inf plus buffer
- In practice, need stopping criterion
- Easy to give stopping criterion if have LB on $\mu_k(R_{\epsilon,M})$ (unrealistic)
- How to do this without knowing a priori essential inf or $R_{\epsilon,M}$?
- Has been shown that even if S compact and convex and $f \in C^2$, each step of alg leaves unsampled square region of nonzero measure, over which f can be redefined so that global min is in unsampled region
- "search for a good stopping criterion seems doomed to fail"

Rates of Convergence

- Measured by distributional characteristics of number of iters or function evals required to reach essential inf (e.g. mean)
- Solis and Wets tested 3 versions of the conceptual alg (1 local search, 2 global search) on various problems (constrained and unconstrained)
- They report results only for

$$\min_{\mathbf{x}\in\mathbb{R}^n} \mathbf{x}'\mathbf{x}$$

  with stopping criterion $\|\mathbf{x}^k\| \leq 10^{-3}$
- Found that mean number of function evals required $\propto n$.

Conclusion and Summary of Part I

- Why use random search techniques?
- How to handle pathological cases? (essential infimum, optimality region)
- Conceptual Algorithm unifies past examples in the literature
- Global and local search methods
- Sufficient conditions for convergence and theorems
- Issue of stopping criteria
- Computational results

Part II: Traditional Sampling Methods

- Transformation method
  - easier to generate Y than X, but well-behaved transformation between the two
- Acceptance-rejection method
  - Generate a RV and subject it to a test (based on a second RV) in order to determine acceptance
- Markov-regression
  - Generate random vector component-wise, using marginal distributions w.r.t. components generated already

Impractical because complexity increases rapidly with dimension.

Part II: Approximate Sampling Methods

- Perform better computationally (efficient)
- generates a sequence of points, whose limiting distribution is equal to target distribution

Hit-and-Run: Generate random point in S, a bounded open subset of $\mathbb{R}^d$, according to some target distribution π.

1. Initialize: select starting point $\mathbf{x}^0 \in S$. $n := 0$.
2. Randomly generate direction θ^n in $\mathbb{R}^d$, according to distribution ν (corresponds to randomly generating a point on a unit sphere).
3. Randomly select step size from $\lambda_n \in \{\lambda \mid \mathbf{x}^n + \lambda\theta_n \in S\}$ according to distribution $L(\mathbf{x}^n, \theta^n)$
4. Set $\mathbf{x}^{n+1} := \mathbf{x}^n + \lambda_n\theta^n$. $n := n + 1$. Repeat.

e.g. generate point according to uniform distribution on S: use all uniform distributions

Further Reading

References

[1] Baba, N., T. Shoman, and Y. Sawaragi. "A Modified Convergence Theorem for a Random Optimization Algorithm," *Information Science*, 13 (1977).

[2] Gaviano, M. "Some General Results on the Convergence of Random Search Algorithms in Minimization Problems." In *Towards Global Optimization*, eds. L. Dixon and G. Szegö. Amsterdam.

[3] Solis, Francisco J. and Roger J.B. Wets. "Minimization by Random Search Techniques," *Mathematics of Operations Research*, 6: 19 - 30 (1981).

[4] H.E. Romeijn, *Global Optimization by Random Walk Sampling Methods*, Thesis Publishers, Amsterdam, 1992.