Stochastic Optimization

Michael W. Trosset
Department of Mathematics
College of William & Mary

trosset@math.wm.edu

http://www.math.wm.edu/~trosset/

Introduction

Consider the problem of minimizing $f : \Re^p \to \Re$, where $f(x)$ must be estimated by observing the value of a random variable $\widehat{f}_n(x)$.

More precisely, let $f(x) = T(P_x)$ and $\widehat{f}_n(x) = T(\widehat{P}_x)$, where $\widehat{P}_x$ is the empirical distribution of an i.i.d. random sample $\omega_1, \ldots, \omega_n \sim P_x$. Then

$$\widehat{f}_n(x) \xrightarrow{P} f(x),$$

so we can obtain more accurate estimates by drawing larger samples.

Such problems routinely arise in simulation-based estimation.

Methods for Numerical Optimization

0. Direct comparison of f values.

   Examples: Grid Search, Pattern Search.

1. Local linear models of f, usually constructed from first derivative information.

   Example: Steepest Descent.

2. Local quadratic models of f, usually constructed from first and second derivative information.

   Example: Newton's Method.

Direct Search

Example: "Compass Search" in $\Re^2$

Do until convergence:

1. From x_c, look Δ units to NSEW.

2. If find $f(x_t) < f(x_c)$, then $x_c \leftarrow x_t$;

   else $\Delta \leftarrow \Delta/2$.

Direct search methods are often recommended for optimization in the presence of random noise, i.e. when function evaluation is uncertain.

Why?

Supporting evidence for this recommendation is largely anecdotal, bolstered by a few simulation studies, e.g.

> R.R. Barton (1984). Minimization algorithms for functions with random noise. *American Journal of Mathematical and Management Sciences*, 4:109–138.

What is invariably considered is how algorithms designed for optimization in the absence of random noise perform in the presence of random noise.

Convergence

To ensure convergence, it is necessary for $n \to \infty$ as $\Delta \to 0$. Recent theoretical results in

- E.J. Anderson & M.C. Ferris (2001), A direct search algorithm for optimization with noisy function evaluation, *SIAM Journal on Optimization*, 11:837–857; and

- M.W. Trosset (2000). On the use of direct search methods for stochastic optimization, Technical Report 00-20, Department of Computational & Applied Mathematics, Rice University, Houston, TX,

suggest that an alarming number of observations are required for direct search to progress as Δ becomes small.

Anderson & Ferris (2001) proposed a class of algorithms for the case that function evaluation is corrupted by adding normal errors:

$$\omega_1, \ldots, \omega_n \sim \text{Normal}\left(f(x), \sigma^2\right)$$

These algorithms employ three operations:

- $\vec{x}_{k+1} = \texttt{reflect}(\vec{x}_k)$, preserving the size of the design, and $n_{k+1} = n_k$;
- $\vec{x}_{k+1} = \texttt{expand}(\vec{x}_k)$, essentially doubling the size of the design, and $n_{k+1} \approx n_k/4$; and
- $\vec{x}_{k+1} = \texttt{contract}(\vec{x}_k)$, essentially halving the size of the design, and $n_{k+1} \approx 4n_k$.

The convergence analysis depends critically on the tail behavior of the normal distribution.

Anderson & Ferris defined the size of a design $\vec{x}$ to be

$$D\left(\vec{x}\right) = \max\left\{\left\|x_i - x_j\right\| : x_i, x_j \in \vec{x}\right\}$$

and decreased $\sigma/\sqrt{n_k}$ faster than $D(\vec{x}_k)$ to ensure convergence. This condition is virtually identical to the condition derived in Trosset's power analysis.

Their numerical results reinforce Trosset's conclusion that using direct search in the presence of random noise is extraordinarily expensive. With $\sigma = 0.1$, they set $\xi(\vec{x}_0) = 1$ and stopped when $\xi(\vec{x}_k) < 0.0001$, where

$$\xi\left(\vec{x}\right) = \min\left\{\left\|x_i - x_j\right\| : x_i, x_j \in \vec{x}\right\}.$$

Thus, after the penultimate contraction, each estimated function value required a sample of size

$$n_k \approx 4^{13} = 67,108,864.$$

Stochastic Approximation

At x_c, choose α^* to minimize

$$g(\alpha) = f\left(x_c - \alpha \nabla f\left(x_c\right)\right);$$

then set

$$x_+ = x_c - \alpha^* \nabla f\left(x_c\right).$$

This is the method of steepest descent.

If derivatives are not available, then they must be approximated, usually by finite differencing:

$$\nabla f\left(x_c\right) \approx \frac{1}{2\beta} \begin{bmatrix} f(x_c + \beta e_1) - f(x_c - \beta e_1) \\ \vdots \\ f(x_c + \beta e_p) - f(x_c - \beta e_p) \end{bmatrix}$$

Both the differences and the line search are highly unstable in the presence of noise.

Classical Stochastic Approximation

J. Kiefer & J. Wolfowitz (1952). Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, 23:462–466.

Let α_k and β_k satisfy

$$\beta_k \to 0, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty,$$

$$\sum_{k=1}^{\infty} \alpha_k \beta_k < \infty, \qquad \sum_{k=1}^{\infty} \alpha_k^2 \beta_k^2 < \infty,$$

e.g. $\alpha_k = k^{-1}$ and $\beta_k = k^{-1/3}$.

The algorithm is

$$x_{k+1} = x_k - \frac{\alpha_k}{2\beta_k} \left[\widehat{f}_n \left(x_k + \beta_k \right) - \widehat{f}_n \left(x_k - \beta_k \right) \right],$$

typically for $n = 1$.

Note: Paul Goger's UMSA project and senior honors thesis explored various strategies for choosing n_k.

Response Surface Methodology

Instead of approximating $\nabla f(x_c)$ by finite differencing. . .

1. Design an experiment, i.e. specify an experimental region E in the vicinity of x_c and choose design sites $w_1, \ldots, w_d \in E$.

2. Perform the experiment, i.e. observe values $y_i = \widehat{f}_n(w_i)$ for $i = 1, \ldots, d$.

3. Fit a local linear model h_1 to the data

$$(w_1, y_1), \ldots, (w_d, y_d),$$

   e.g. by least-squares regression.

4. Compute $\nabla h_1(x_c)$ and proceed as with stochastic approximation.

Quadratic Models

For numerical optimization, we often write

$$\begin{aligned} f(x) &\approx f(x_c) + (x - x_c)^T \nabla f(x_c) + \\ &\quad \frac{1}{2}(x - x_c)^T \left[\nabla^2 f(x_c)\right](x - x_c) \\ &= q(x) \end{aligned}$$

and set x_+ equal to the stationary point of the quadratic approximation q. This is equivalent to solving $\nabla f(x) = \vec{0}$ by Newton's method.

Alternatively, we might choose x_+ to solve

$$\begin{aligned} &\text{minimize} && q(x) \\ &\text{subj to} && \|x - x_c\| \leq r_c, \end{aligned}$$

where the constraint set is the region in which we "trust" the quadratic approximation q.

When function evaluation is uncertain. . .

1. Design an experiment, i.e. specify an experimental region E in the vicinity of x_c and choose design sites $w_1, \ldots, w_d \in E$.

2. Perform the experiment, i.e. observe values $y_i = \widehat{f}_n(w_i)$ for $i = 1, \ldots, d$.

3. Fit a local quadratic model h_2 to the data

$$(w_1, y_1), \ldots, (w_d, y_d),$$

   e.g. by least-squares regression, and proceed as above.

Remark: In RSM, solving the trust region subproblem is called *ridge analysis*.