

# Quasi-Newton Methods for Stochastic Optimization With Application to Simulation-Based Parameter Estimation

Brent Castle <sup>a,b</sup>

Michael Trosset <sup>b,a</sup>

This work was supported by a grant from the Air Force Office of Scientific Research.

a - School of Informatics and Computing, Indiana University

b - Department of Statistics, Indiana University

# Outline

1. Stochastic Optimization and Simulated-Based Parameter Estimation
2. Motivating Example - Tumor Recurrence
3. QNSTOP - Quasi-Newton Methods for Stochastic Optimization
4. Illustrative Example - Fitting a two parameter stochastic process
5. Numerical Experiments - Tumor Recurrence
6. Future Work

# Stochastic Optimization

Let  $\mathcal{P} = \{P(\cdot; \theta) : \theta \in \Theta \subset \mathbb{R}^p\}$  denote a family of probability distributions.

Let  $\omega_1, \dots, \omega_n \sim P(\cdot; \theta)$  be an IID sample. The empirical probability distribution of the sample is

$$\hat{P}_n(x; \theta) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(\omega_i \leq x).$$

Let  $T : \mathcal{P}^* \rightarrow \mathbb{R}$  denote a statistical functional where its domain,  $\mathcal{P}^*$ , contains all possible empirical distributions  $\hat{P}_n$  for samples ( $n \geq 1$ ) drawn from  $P \in \mathcal{P}$  as well as  $P$  itself.

As  $n \rightarrow \infty$ ,

$$\sqrt{n} \left[ T(\hat{P}_n) - T(P) \right] \rightarrow \mathcal{N}(0, \tau^2)^1$$

where  $\tau^2 = \int L_P(x)^2 dP(x)$ .

---

<sup>1</sup>See Fernholz (1983). von Mises Calculus for Statistical Functionals.

# Stochastic Optimization

We seek to minimize  $f(\theta) = T(P(\cdot; \theta))$  subject to  $\theta \in \Theta$ .

If  $P$  is unknown or analytically intractable then minimizing  $f(\theta)$  is a stochastic optimization problem.

We obtain information about  $P(\cdot; \theta)$  by drawing  $\omega_1, \dots, \omega_n \sim P(\cdot; \theta)$  and estimating

$$f(\theta) = T(P(\cdot; \theta))$$

by

$$\hat{f}_n(\theta) = T(\hat{P}_n(\cdot; \theta)).$$

# Simulation-Based Parameter Estimation

Let  $q_1, \dots, q_m \sim Q$  be an observed sample where  $Q$  is an unknown distribution we estimate by the empirical distribution  $\widehat{Q}_m$ . We seek to estimate  $\theta$  such that  $P(\cdot; \theta)$  estimates  $Q$ .

Let the functional  $\Delta(\cdot, \cdot)$  measure the discrepancy between two distributions. We would like to estimate  $\theta$  by minimizing

$$f(\theta) = T(P(\cdot; \theta)) = \Delta \left( P(\cdot; \theta), \widehat{Q}_m \right),$$

but evaluation of  $f$  is intractable. Instead, we estimate  $f(\theta)$  with

$$\widehat{f}(\theta) = T \left( \widehat{P}_n(\cdot; \theta) \right) = \Delta \left( \widehat{P}_n(\cdot; \theta), \widehat{Q}_m \right),$$

where  $\widehat{P}_n$  is the empirical distribution of a simulated sample. With this substitution, the problem of minimum distance estimation becomes a problem of stochastic optimization.

# Minimum Distance Estimation

The following are two possible choices for  $\Delta$ .

## 1. Two-sample Kolmogorov-Smirnov Statistic

$$\begin{aligned}\hat{f}_{\text{K-S}}(\theta) &= \Delta \left( \hat{P}_n(\cdot; \theta), \hat{Q}_m \right) \\ &= \sup_x \left| \hat{P}_n(x; \theta) - \hat{Q}_m(x) \right|\end{aligned}$$

## 2. Anderson's Statistic for the Two-Sample Cramér-von Mises Test

$$\begin{aligned}\hat{f}_{\text{C-vM}}(\theta) &= \Delta \left( \hat{P}_n(\cdot; \theta), \hat{Q}_m \right) \\ &= \frac{nm}{n+m} \int_{-\infty}^{\infty} \left( \hat{P}_n(x; \theta) - \hat{Q}_m(x) \right)^2 d\hat{H}_{n+m}(x)\end{aligned}$$

where  $(n+m)\hat{H}_{n+m} = n\hat{P}_n + m\hat{Q}_m$ .

## Motivating Example

Atkinson, Bartoszynski, Brown, and Thompson (1983) model tumor recurrence, i.e., the time between detection of a primary and secondary tumor, by the following axioms:

## Motivating Example

Atkinson, Bartoszynski, Brown, and Thompson (1983) model tumor recurrence, i.e., the time between detection of a primary and secondary tumor, by the following axioms:

1. Each tumor originates from a single cell and grows exponentially at rate  $\theta_1$ .

# Motivating Example

Atkinson, Bartoszynski, Brown, and Thompson (1983) model tumor recurrence, i.e., the time between detection of a primary and secondary tumor, by the following axioms:

1. Each tumor originates from a single cell and grows exponentially at rate  $\theta_1$ .
2. Occurrence of systemic tumors is a Poisson process with rate  $\theta_2$ .

# Motivating Example

Atkinson, Bartoszynski, Brown, and Thompson (1983) model tumor recurrence, i.e., the time between detection of a primary and secondary tumor, by the following axioms:

1. Each tumor originates from a single cell and grows exponentially at rate  $\theta_1$ .
2. Occurrence of systemic tumors is a Poisson process with rate  $\theta_2$ .
3. Detection of tumor  $j$  is a nonhomogeneous Poisson process with rate  $\theta_3 Y_j(t)$ , where  $Y_j(t)$  is the size of tumor  $j$  at time  $t$ .

## Motivating Example

Atkinson, Bartoszynski, Brown, and Thompson (1983) model tumor recurrence, i.e., the time between detection of a primary and secondary tumor, by the following axioms:

1. Each tumor originates from a single cell and grows exponentially at rate  $\theta_1$ .
2. Occurrence of systemic tumors is a Poisson process with rate  $\theta_2$ .
3. Detection of tumor  $j$  is a nonhomogeneous Poisson process with rate  $\theta_3 Y_j(t)$ , where  $Y_j(t)$  is the size of tumor  $j$  at time  $t$ .
4. Until the removal of the primary tumor, metastasis is a nonhomogeneous Poisson process with rate  $\theta_4 Y_0(t)$ .

## Motivating Example

Let  $\text{Time} \sim P(\cdot; \theta)$  denote the time from detection of the first tumor to detection of the second tumor.  $P(\cdot; \theta)$  is intractable, but easily sampled by stochastic simulation:

```
Repeat until Time >  $\theta$ 
  Generate  $U_1, U_2, U_3, U_4 \sim \text{Unif}(0, 1)$ 
  Detect1  $\leftarrow \log(1 - (\theta_1/\theta_3) \log U_1)/\theta_1$ 
  Metastasis  $\leftarrow \log(1 - (\theta_1/\theta_4) \log U_2)/\theta_1$ 
  NewSystemic  $\leftarrow (-\log U_3)/\theta_2$ 
  If Metastasis > Detect1 then
    Second  $\leftarrow \text{NewSystemic}$ 
  Else
    Second  $\leftarrow \min(\text{Metastasis}, \text{NewSystemic})$ 
  Detect2  $\leftarrow \log(1 - (\theta_1/\theta_3) \log U_4)/\theta_1$ 
  Time  $\leftarrow \text{Second} + \text{Detect2} - \text{Detect1}$ 
```

## Motivating Example

Time was observed for  $m = 116$  breast cancer patients at the Curie-Sklodowska Cancer Institute in Warsaw.

Let  $\widehat{Q}_m$  denote the empirical distribution of these times.

The minimum distance estimator  $\hat{\theta} \in \Theta$  is defined as the value that

$$\Delta \left( P(\cdot; \hat{\theta}), \widehat{Q}_m \right) = \inf_{\theta \in \Theta} \Delta \left( P(\cdot; \theta), \widehat{Q}_m \right)$$

which we seek to estimate by minimizing

$$\hat{f}(\theta) = \Delta \left( \widehat{P}_n(\cdot; \theta), \widehat{Q}_m \right).$$

# Response Surface Methodology

Box and Wilson (1951) introduced Response Surface Methodology (RSM) as a means of exploring the response surface.

1. Constructs a sequence of designed regression experiments.
2. Linear and quadratic approximations of  $f$ .
3. Research emphasizes experimental design, but not automation or convergence theory.

# Quasi-Newton Methods for Stochastic Optimization

QNSTOP progresses by constructing ellipsoids

$$E_k(\eta) = \left\{ x \in \mathbb{R}^p : (x - \hat{\xi}_k)^t W_k (x - \hat{\xi}_k) \leq \eta^2 \right\}$$

by the following five steps.

1. A space filling design  $(\{x_i\}_{i=1}^{N_k})$  is constructed in  $E_k(\tau_k)$ .
  - 1.1 Sample  $M \gg N_k$  design sites uniformly in  $E_k(\tau_k)$ .
  - 1.2 Remove design sites with small pairwise distances to leave a space filling design.
2. We observe  $\hat{f}(x) = \Delta(\hat{P}_n(\cdot; x), \hat{Q}_m)$  for each design site.

# Quasi-Newton Methods for Stochastic Optimization

3. Let  $s_i = x_i - \hat{\xi}_k$ .

(a) If enough design sites are available in  $E_k$  a quadratic model  $\hat{m}_k(s)$  is fit to the observations by least-squares regression.

(b) Otherwise, we fit a linear model and update the Hessian by the BFGS update.

Let  $g_k$  be the estimate of the gradient,  $\lambda_k = g_k - g_{k-1}$ , and  $\nu_k = \hat{\xi}_k - \hat{\xi}_{k-1}$ . We update the approximate Hessian by

$$H_k = H_{k-1} + \frac{\lambda_k \lambda_k^T}{\nu_k^T \lambda_k} - \frac{H_{k-1} \nu_k \nu_k^T H_{k-1}}{\nu_k^T H_{k-1} \nu_k}.$$

# Quasi-Newton Methods for Stochastic Optimization

4. Our new estimate is the minimizer of the quadratic model subject to a trust region constraint, i.e.,

$$\begin{aligned} \min \quad & \hat{m}_k(s) = f_k + g_k^t s + \frac{1}{2} s^t H_k s \\ \text{s. t.} \quad & \hat{\xi}_k + s \in E_k(\rho_k) \end{aligned}$$

and set  $\hat{\xi}_{k+1} = \hat{\xi}_k + \hat{s}$ .

The step

$$\hat{s} = -[H_k + \hat{\mu}_k W_k]^{-1} g_k$$

is solved by computing  $\hat{\mu}_k$ , the Lagrange multiplier of the trust region subproblem.<sup>2</sup>

---

<sup>2</sup>See chapter 7 in Conn, Gould, and Toint. Trust Region Methods, 2000.

# Quasi-Newton Methods for Stochastic Optimization

5. We compute the covariance (or an approximation),  $V$ , of  $\nabla m_k(\hat{s}) = g_k + H_k \hat{s}$ .

Let  $\delta(s) = [H_k + \hat{\mu}_k W_k]s + g_k$ , an estimated stationary point of the Lagrangian. The set of  $s$  that satisfy

$$\delta(s)^T V^{-1} \delta(s) \leq \chi_{p,1-\alpha}$$

is an approximation of the  $1 - \alpha$  percentile confidence set of the minimizer of the quadratic subject to the trust region constraint.<sup>3</sup>

The ellipsoid shape parameter is then updated by

$$W_{k+1} = (H_k + \hat{\mu}_k W_k)^t V^{-1} (H_k + \hat{\mu}_k W_k)$$

and subsequently its eigenvalues are modified to prevent it from collapsing and it is scaled so its determinant is 1.

---

<sup>3</sup>Stablein et. al. Confidence regions for constrained optima in response-surface experiments. Biometrics, 1983.

## Illustrative Example

We applied QNSTOP to the problem of estimating the parameters (drift ( $\mu$ ) and volatility( $\sigma$ )) of a geometric Brownian motion from an observed sample of values at a fixed time. Recall that a stochastic process  $S_t$  is a Geometric Brownian Motion if it is described by the SDE

$$dS_t = \mu S_t dt + \sigma S_t dW_t$$

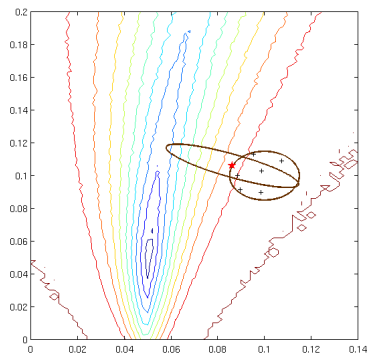
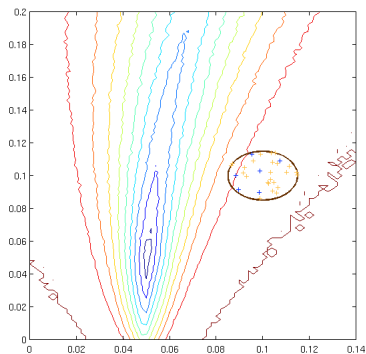
where  $W_t$  is a Wiener process.

A sample ( $m = 500$ ) was generated by observing a GBM ( $\mu = 0.05, \sigma = 0.05$ ) at  $t = 50$ . The empirical distribution of the sample was computed ( $\hat{Q}$ ) and treated as a reference sample.

For each observation of  $\hat{f}(x)$  a sample ( $n = 500$ ) was generated to compute  $\hat{P}_n$ .

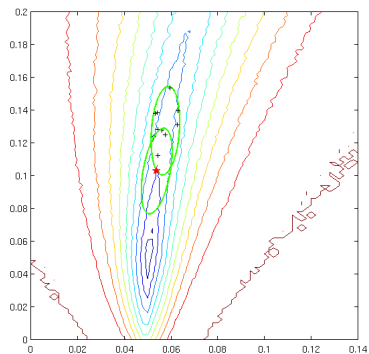
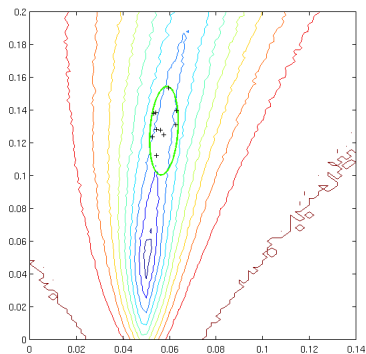
# Illustrative Example

## Iteration 1



# Illustrative Example

Iteration 6



# Tumor Recurrence

Recall the tumor recurrence model from Atkinson et al. (1983).

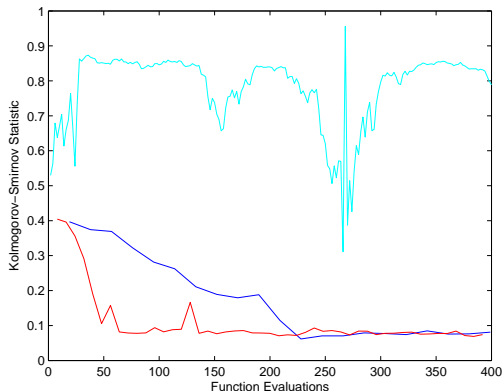
We generated a sample ( $m = 500$ ) to represent actual tumor recurrence data, but at a known  $\theta$ .

Let  $\hat{P}_n(\cdot; \theta)$  denote the empirical distribution of our sample ( $n = 5000$ ) generated for each function evaluation ( $\hat{f}_n(\theta)$ ).

We estimate  $\theta$  with two versions of QNSTOP (Full Quadratic and BFGS) and Simultaneous Perturbation Stochastic Approximation (Spall, 1992). Each is given a budget of 400 function evaluations.

# Tumor Recurrence

Kolmogorov-Smirnov



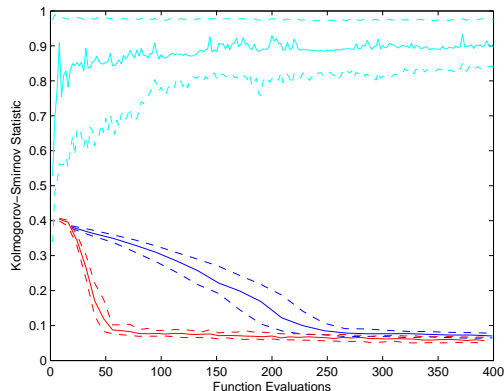
SPSA -  $N_k = 2$ .

QNSTOP (Full Quadratic) -  $N_k = 19$ .

QNSTOP (BFGS) -  $N_k = 8$ .

# Tumor Recurrence

Kolmogorov-Smirnov - Quartiles from 100 repetitions.



SPSA -  $N_k = 2$ .

QNSTOP (Full Quadratic) -  $N_k = 19$ .

QNSTOP (BFGS) -  $N_k = 8$ .

## Future Work

- ▶ Considerations for  $\hat{\xi}_{k+1} = \Pi_{\Theta}(\hat{\xi}_k + \hat{s})$ .
- ▶ Convergence theory for unconstrained and constrained algorithms.
- ▶ Application to “non-smooth” deterministic functions with expensive function evaluations.

# References

Fernholz. *von Mises calculus for statistical functionals*. Springer-Verlag, 1983.

Atkinson, Bartoszynski, Brown, and Thompson. *Simulation techniques for parameter estimation in tumor related stochastic processes*. Proceedings of the 1983 Computer Simulation Conference.

Box and Wilson. *On the experimental attainment of optimum conditions*. JRSS, Series B, 1951.

Conn, Gould, and Toint. *Trust Region Methods*. 2000.

Stablein, Carter, Jr., and Wampler. *Confidence regions for constrained optima in response-surface experiments*. Biometrics, 1983.

Spall. *Multivariate stochastic approximation using a simultaneous perturbation gradient approximation*. IEEE Transactions on Automatic Control, 1992.

Spall. *Implementation of the simultaneous perturbation algorithm for stochastic optimization*. IEEE Transactions on Aerospace and Electronic Systems, 1998.