Statistical Quasi-Newton:

A New Look at Least-Change

Scott Vander Wiel and Chuanhai Liu

Statistics and Data Mining Research
Bell Labs, Lucent Technologies
Murray Hill, NJ 07974

Paper: stat.bell-labs.com/scottyv

Notation Reference Sheet

x	point in $\mathcal{R}^n$	x, x_+	current, next iterate
$f(x)$	function to minimize	$\tilde{x}$	canonical coordinates
$g(x)$	gradient of f	$\tilde{f}, \tilde{g}$	canonical f, g
$H(x)$	Hessian of f	$\tilde{H}, \tilde{B}$	canonical H, B
B	estimate of H		
$\mathcal{M}^+$	symmetric p.d. matrices		

$$\tilde{B}_+ = \begin{bmatrix} a & b' \\ b & C \end{bmatrix}$$

d	$\equiv -B^{-1}g$	quasi-Newton step direction
s		step size
δ	$\equiv x_+ - x = sd$	step increment
γ	$\equiv g_+ - g$	gradient increment
ϕ		Broyden parameter
λ	$= 1 + \phi/a$	alternative Broyden parameter

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

where $f(x)$ and $g(x) \equiv \nabla f(x)$ are easy to compute.

However, the Hessian $H(x) \equiv \nabla^2 f(x)$ is not.

Initialize: $B \in \mathcal{M}^+$, $x \in \mathcal{R}^n$, $g = g(x)$

Minimize: Search in quasi-Newton direction

$$d \equiv -B^{-1}g$$

for step size $s > 0$ to obtain

$$x_+ = x + sd \quad \text{and} \quad g_+ = \nabla f(x_+),$$

satisfying *Sufficient Decrease* & *Curvature* conditions

Estimate: Update approximate Hessian

$$B_+ = \texttt{update}(B, x, x_+, g, g_+) \in \mathcal{M}^+$$

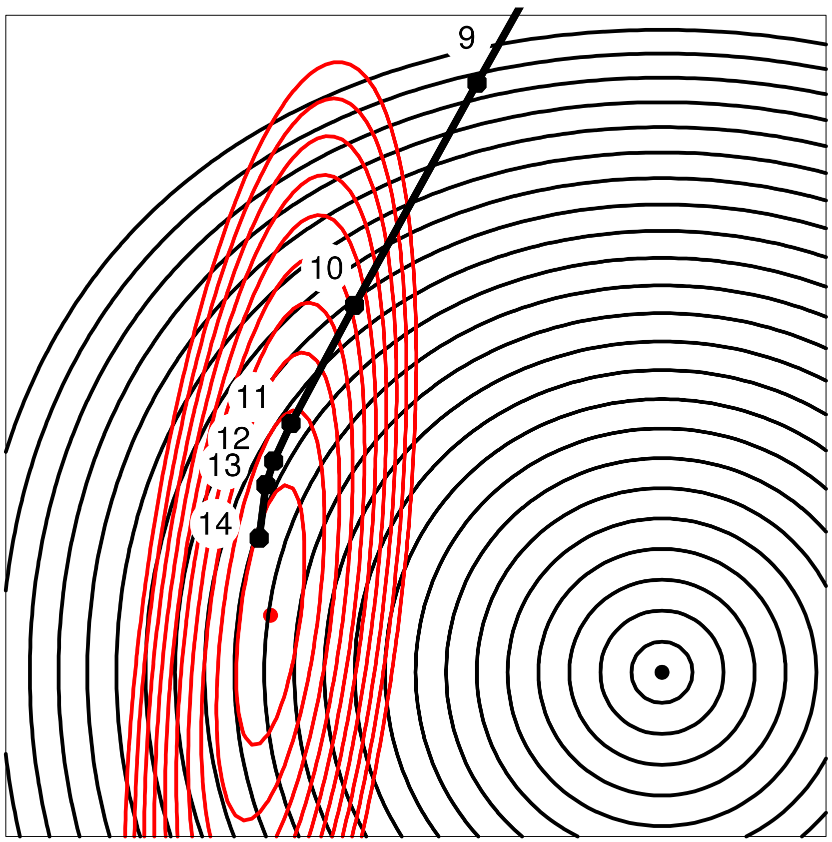
with *quasi-Newton condition*

$$B_+\delta = \gamma$$

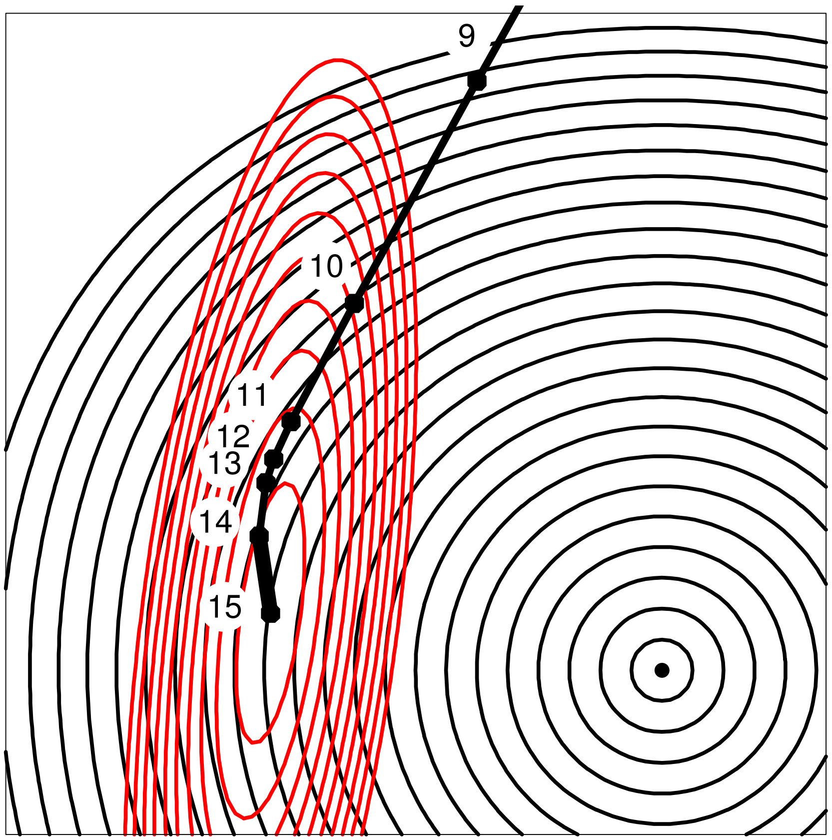
where

$$\delta \equiv x_+ - x \quad \text{and} \quad \gamma \equiv g_+ - g.$$

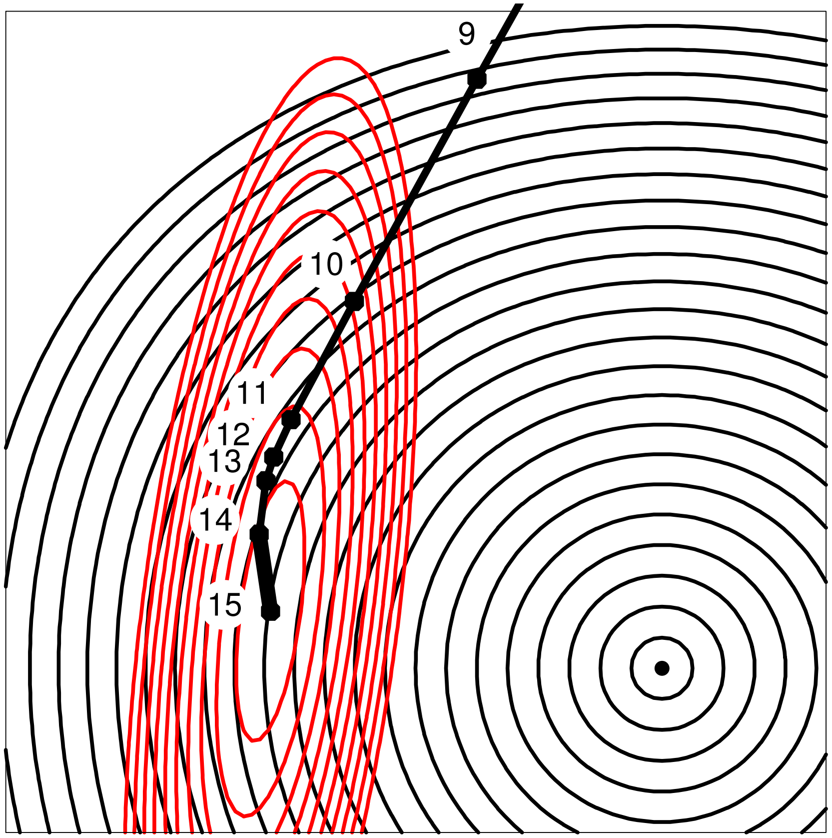
true function (H) **quadratic model (B)**



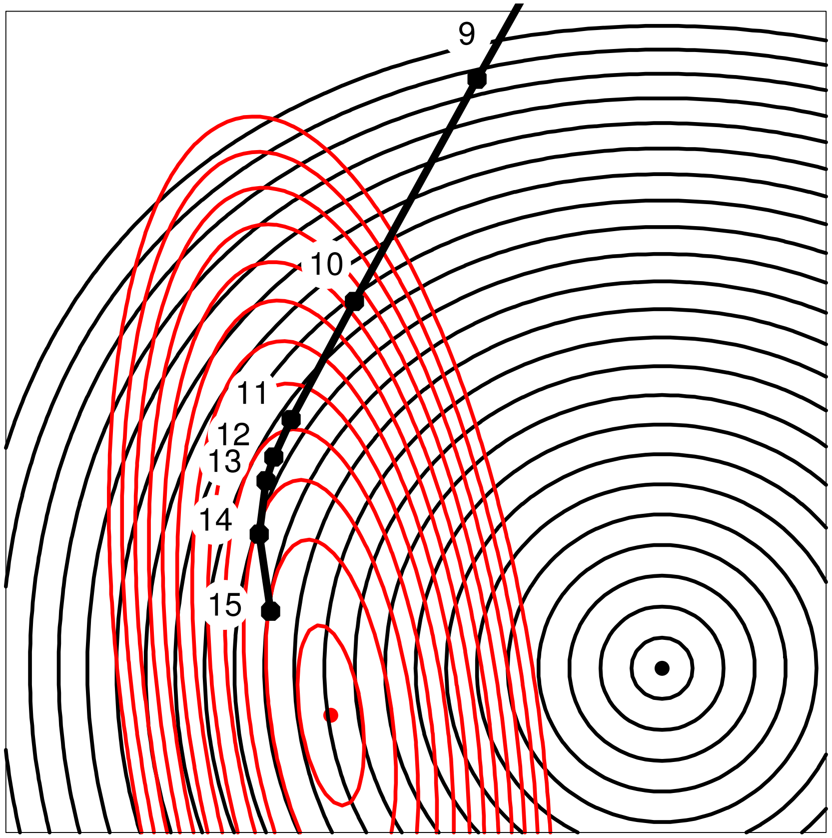
Minimize

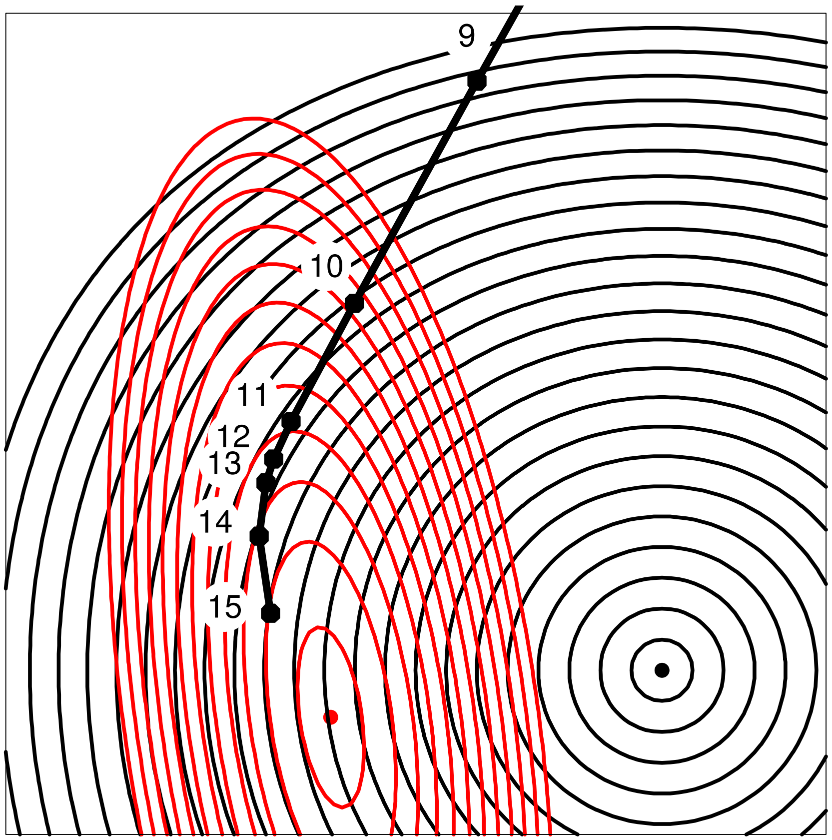


Minimize

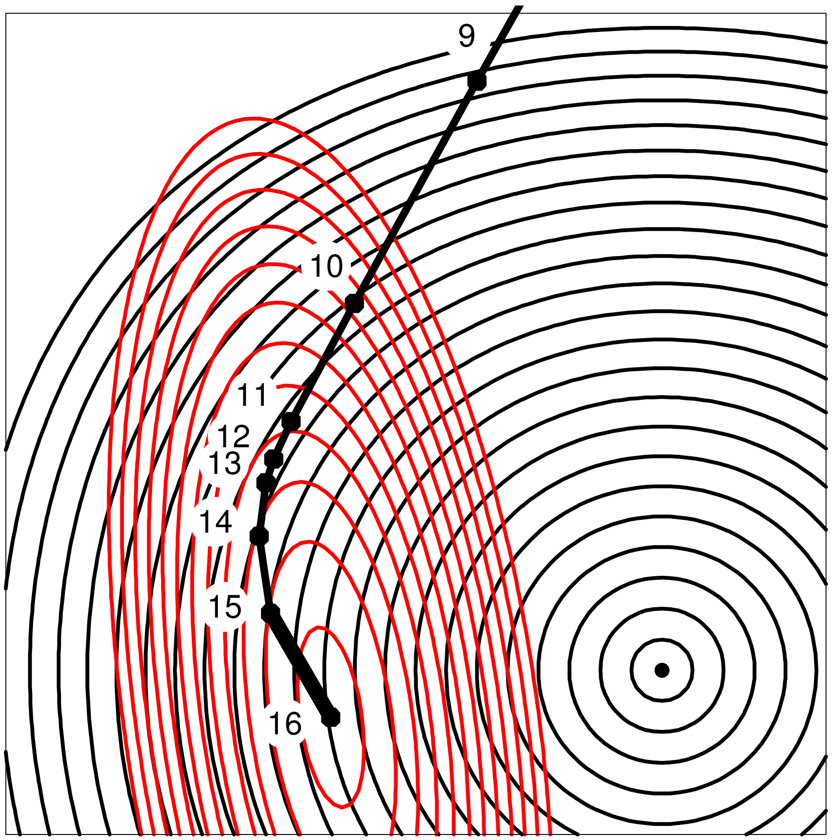


Estimate

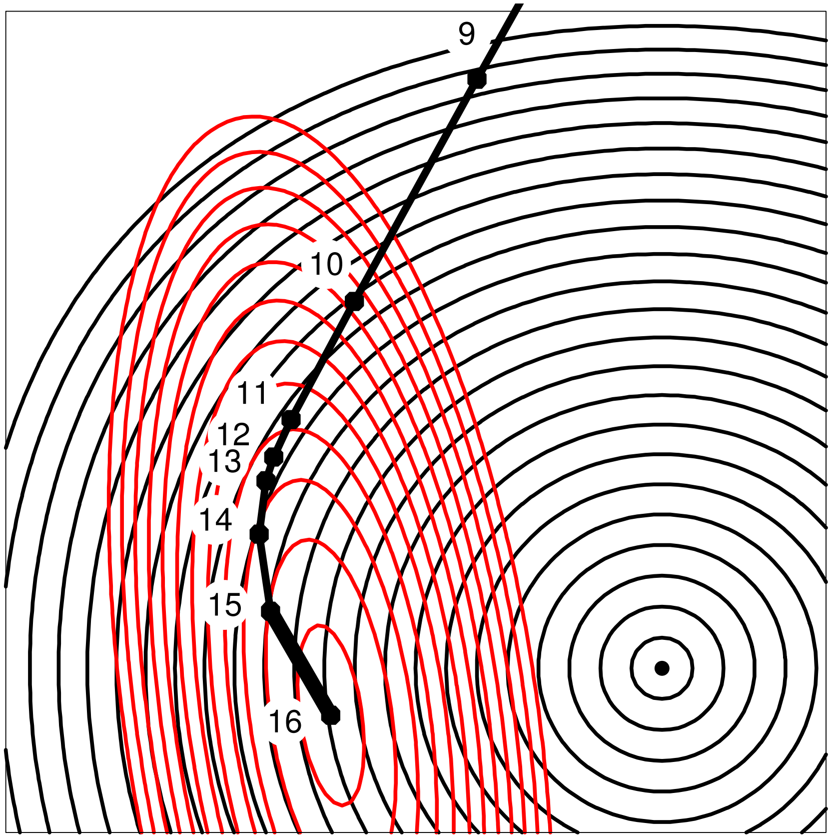



9
10
11
12
13
14
15

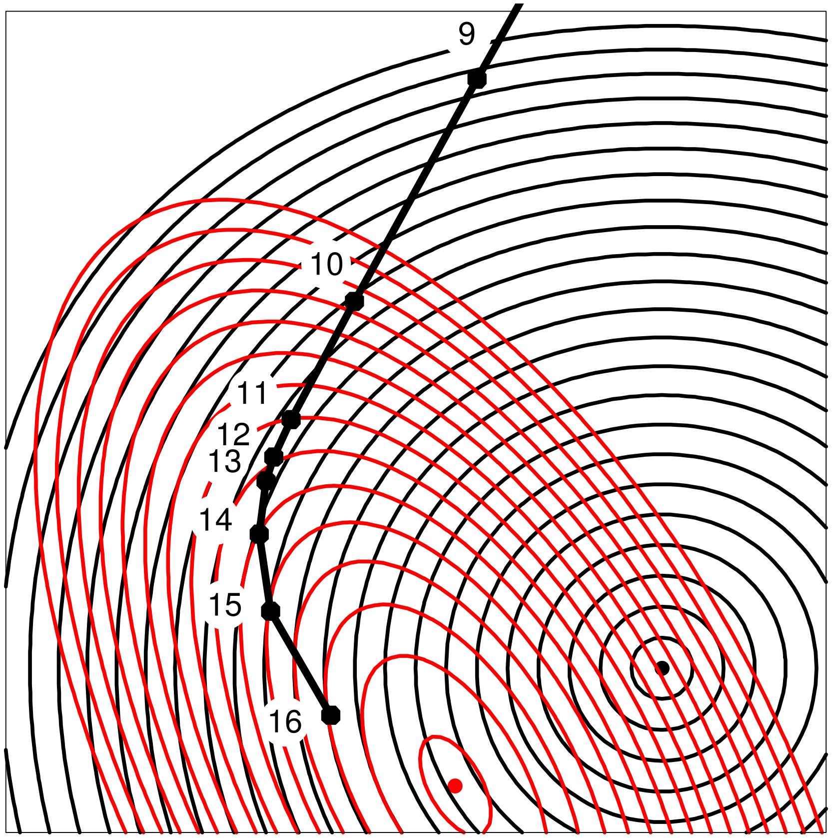
Minimize



Minimize



Estimate



Broyden (1965) — (2000, On the discovery of the "good Broyden" method):

> *We should therefore require, if possible, ...,* ***no change*** *to B in any direction orthogonal to δ.*

Broyden (1967): moved from "no change" principle to ...

> *Since a matrix B^{-1} which possesses to some extent the properties of the inverse Jacobian matrix is already available it would appear reasonable to obtain B_+^{-1} by adding some correction to B^{-1} ...*
>
> $$B_+^{-1} = B^{-1} + C.$$

Led to the Broyden class of rank-2 updates

Broyden Family — symmetric rank-2 updates (Broyden, 1967):

$$B_+(\phi) = B - \frac{B\delta\delta' B}{\delta' B\delta} + \frac{\gamma\gamma'}{\delta'\gamma} + \phi\,(\delta' B\delta) w w',$$

where

$$w \equiv \frac{\gamma}{\delta'\gamma} - \frac{B\delta}{\delta B\delta}$$

ϕ (Broyden parameter)

1 — DFP (Davidon, 1959; Fletcher and Powell, 1963)

0 — BFGS (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970)

Zhang and Tewarson (1988)

Byrd, Liu, and Nocedal (1992)

Fletcher (1994)

ϕ^c — critical value for $B_+(\phi) \in \mathcal{M}^+$

Summary

1. Most popular methods derive from a *Least-Change* principle

$$\min_{\substack{B_+ \in \mathcal{M}^+ \\ B_+\delta = \gamma}} ||B_+ - B||$$

for *some* matrix norm.

But why use Least-Change?

2. Negative Broyden parameters are promising ... but

   - *"Investigations have not shaken BFGS as ... front-runner"* (Zhang and Tewarson, 1988)
   - quasi-Newton steps are often too long

3. Negative Broyden parameters remain mysterious

   - How to choose negative Broyden parameters?
   - How to estimate the step size?

Linear Transformation: Normalize and Rotate :

$$\tilde{x} = U' B^{1/2} x,$$

where U is orthonormal and $U[,1] \propto B^{-1/2} g$.

Transformed Hessian Estimate : $\tilde{B} = I$.

Search direction : $(1, 0, ..., 0)'$, the first axis.

New information on $\tilde{B}_+$: numerical second derivative along $(1, 0, ..., 0)'$

$$\begin{bmatrix} a \\ b \end{bmatrix} \equiv \frac{\tilde{g}_+ - \tilde{g}}{\tilde{x}_+[1] - \tilde{x}[1]}$$

with scalar $a > 0$.

Requirement:

$$B_+\delta = \gamma, \quad \text{and} \quad B_+ \in \mathcal{M}^+$$

Equivalent Requirement:

$$\tilde{B}_+ = \begin{bmatrix} a & b' \\ b & C \end{bmatrix},$$

for C such that

$$C - bb'/a \in \mathcal{M}^+.$$

How to estimate C?

A Naive Update: $C = I$, no change!

$$\tilde{B} = \begin{bmatrix} 1 & 0' \\ 0 & I \end{bmatrix} \quad \Longrightarrow \quad \tilde{B}_+ = \begin{bmatrix} a & b' \\ b & I \end{bmatrix}$$

Why No Change?

– Previous iterations $\Rightarrow$ I is accurate in some directions.

– Updates should not degrade accuracy. Only “no change” always preserves accuracy

– Future iterations will improve poorly estimated directions.

What to do if $a \leq b'b$?

Theorem: The solution to $\min_{\substack{\tilde{B}_+ \in \mathcal{M}^+ \\ \tilde{B}_+ \delta = \gamma}} ||\tilde{B}_+ - I||_{\text{Frobenius}}$

is

$$\tilde{B}_+ = \begin{bmatrix} a & b' \\ b & I + \lambda \frac{bb'}{a} \end{bmatrix} \qquad (*)$$

with

$$\lambda = \lambda_{\text{SQN}} \equiv \begin{cases} 0 & \text{if } a > b'b \\ 1 - r^{-1} & \text{otherwise} \end{cases} \quad \text{and} \quad r \equiv b'b/a.$$

Broyden Family: Has (canonical) form $(*)$ with

$$\lambda = 1 + \phi/a.$$

Thus $\lambda_{\text{BFGS}} = 1$ and λ_{SQN} is a *negative Broyden* update ($\phi < 0$).

Step Sizes

Inexact line-search: Search in direction $d = -B_+^{-1} g_+$ for a step size s that results in sufficient progress.

Initial step size. $s^0 = 1$ is the usual trial step but unit steps are often too large for negative Broyden parameters. (Our tests confirm Zhang and Tewarson (1988).)

Can we estimate the step size?

Wishart Model

Wishart Model: for $\tilde{B}^{+}$

$$\nu \begin{bmatrix} a & b' \\ b & C \end{bmatrix} \sim \text{Wishart}_n\left(\begin{bmatrix} 1 & 0 \\ 0 & I_{n-1} \end{bmatrix}, \nu\right)$$

ν is the degrees of freedom

Use this for estimating initial step size.

Sidebar: Wishart $\Rightarrow$ BFGS

$$\mathsf{E}\left(C \mid a, b\right) = \frac{\nu - 1}{\nu} I + \lambda_{\mathsf{BFGS}} \frac{bb'}{a}.$$

Taking $\nu \to \infty$ gives BFGS!

Step sizes from Wishart

On quadratic functions: optimal step size is

$$s(\lambda) = \frac{d'_+ B_+ d_+}{d'_+ H(x_+) d_+},$$

where $H(x_+)$ is the unknown Hessian.

Wishart model gives:

$$\hat{s} \equiv \lim_{\nu \to \infty} \mathsf{E}\left(s \middle| a, b\right) = \frac{1}{1 + (1-\lambda)\tau_\lambda}$$

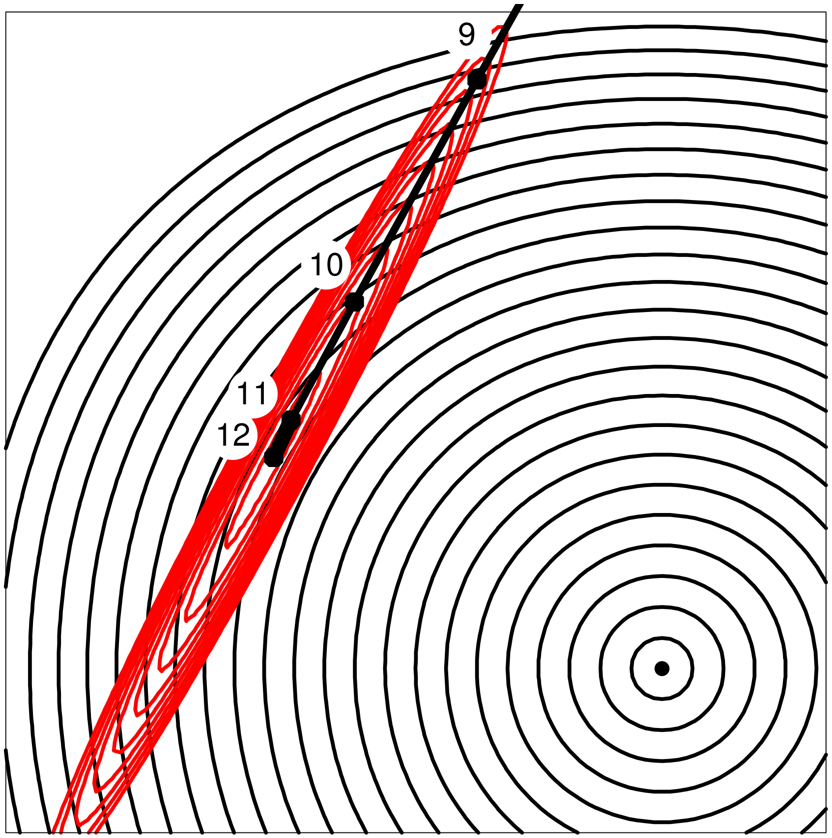
where $\tau_\lambda \equiv \delta' \gamma \left(\omega' d_+\right)^2 / (d'_+ B_+ d_+) \geq 0.$

Special cases:

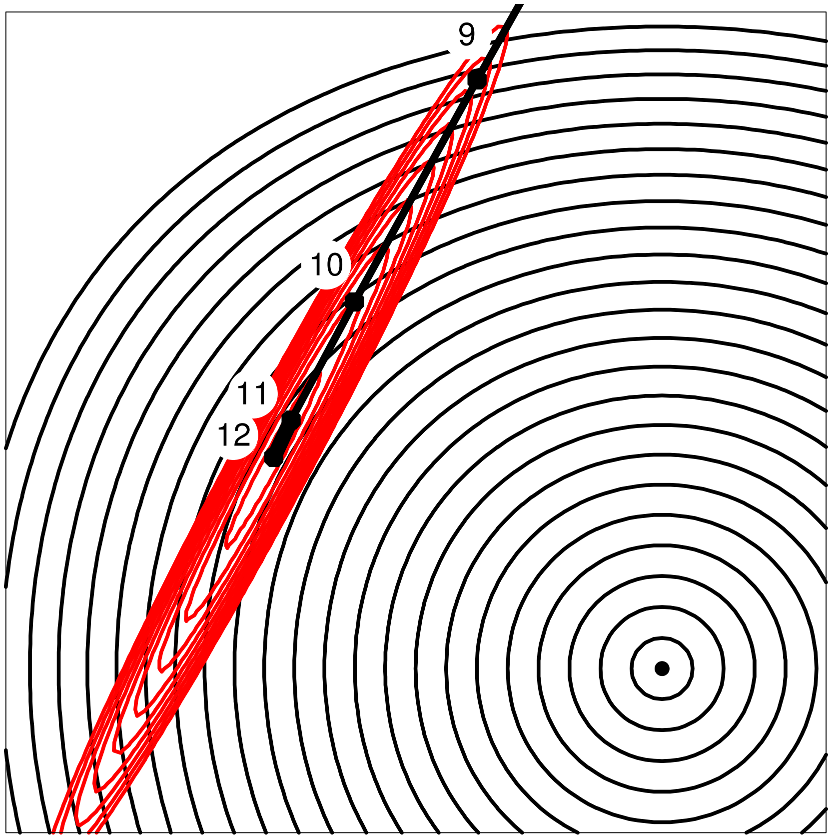
BFGS: $\hat{s} = 1$

SQN: $\hat{s} = (1 + \tau_0)^{-1} \leq 1$

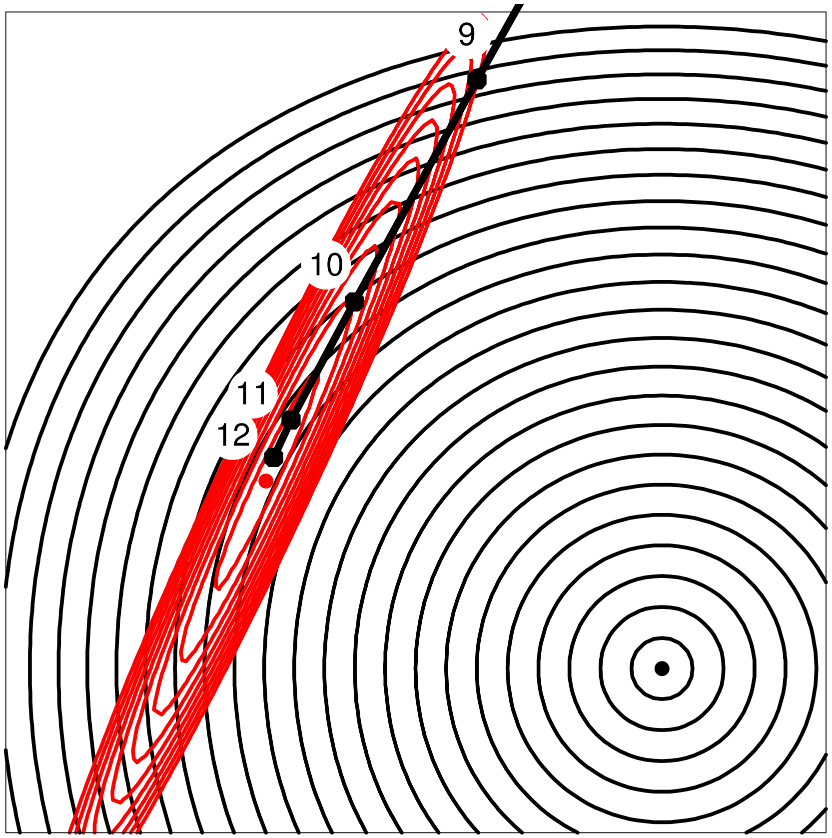
BFGS $(\lambda = 1, \hat{s} = 1)$



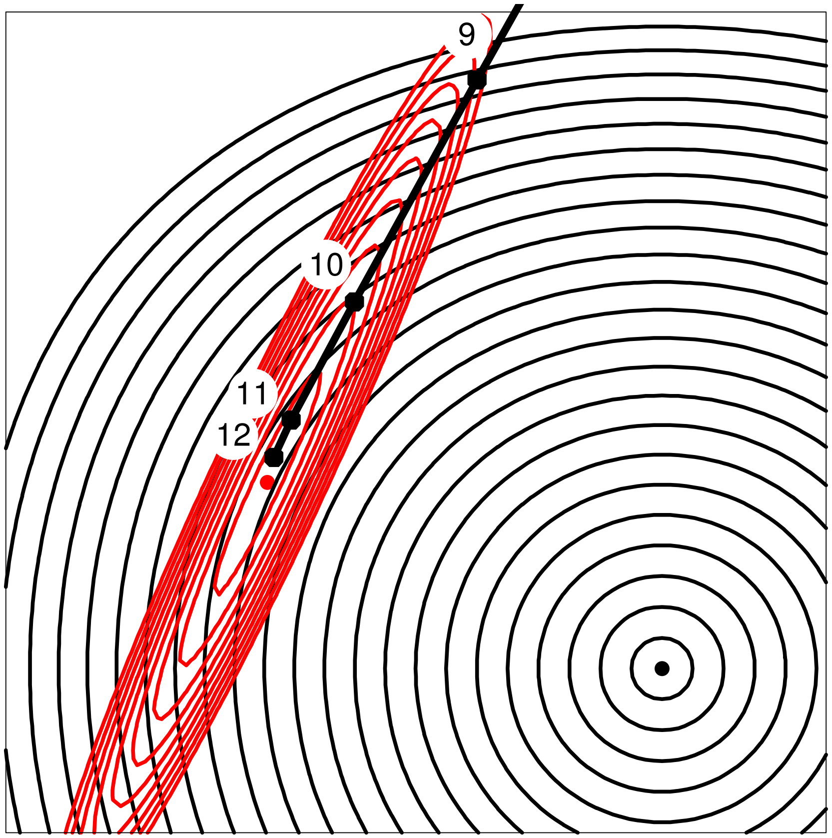
SQN $(\lambda = 0^+, \hat{s} \leq 1)$



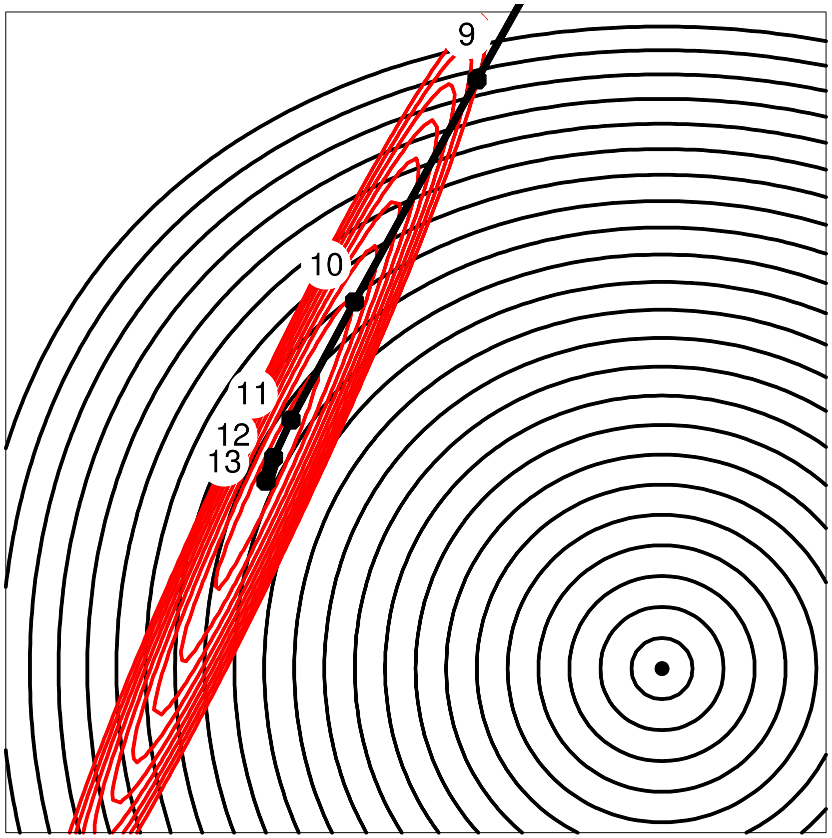
BFGS $(\lambda = 1, \hat{s} = 1)$



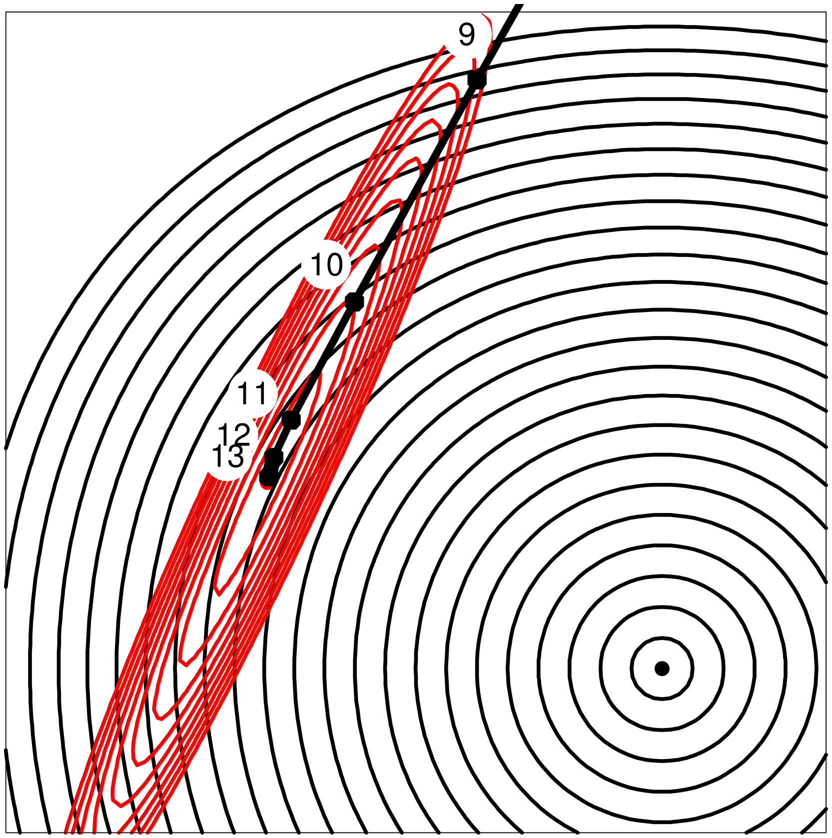
SQN $(\lambda = 0^{+}, \hat{s} \leq 1)$



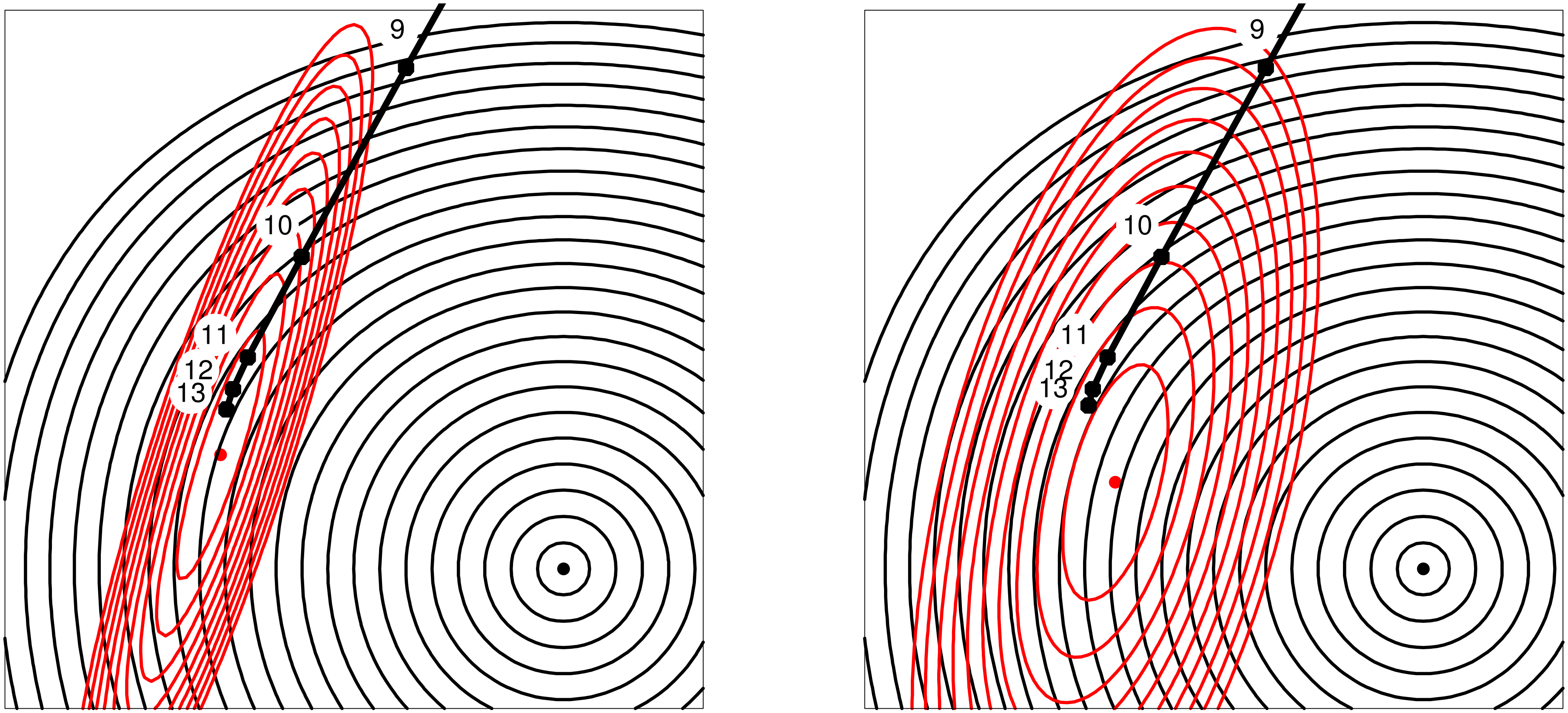
BFGS $(\lambda = 1,\ \hat{s} = 1)$



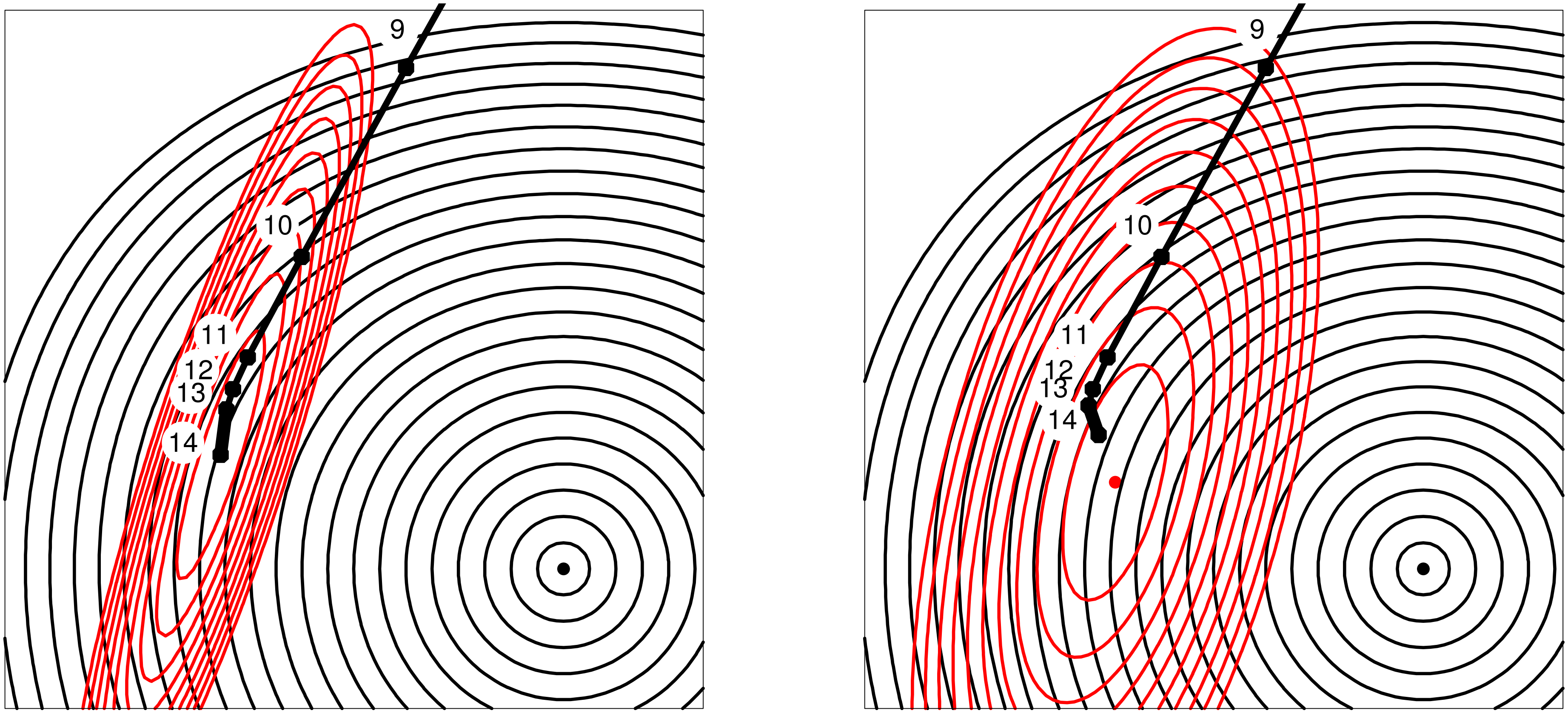
SQN $(\lambda = 0^+,\ \hat{s} \leq 1)$



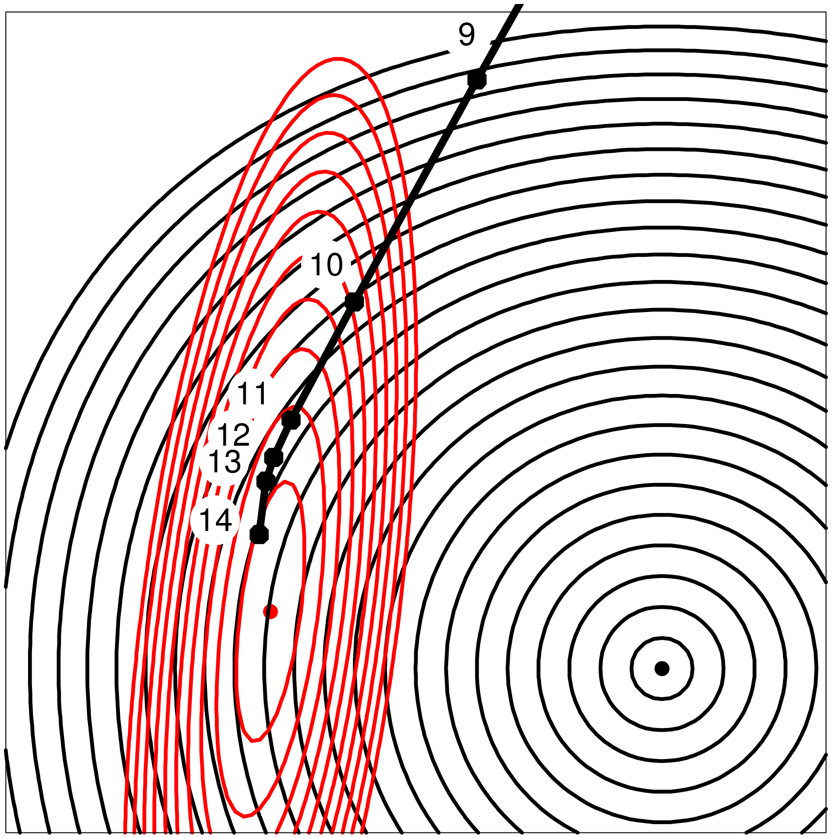
BFGS $(\lambda = 1,\ \hat{s} = 1)$ **SQN** $(\lambda = 0^+, \hat{s} \leq 1)$



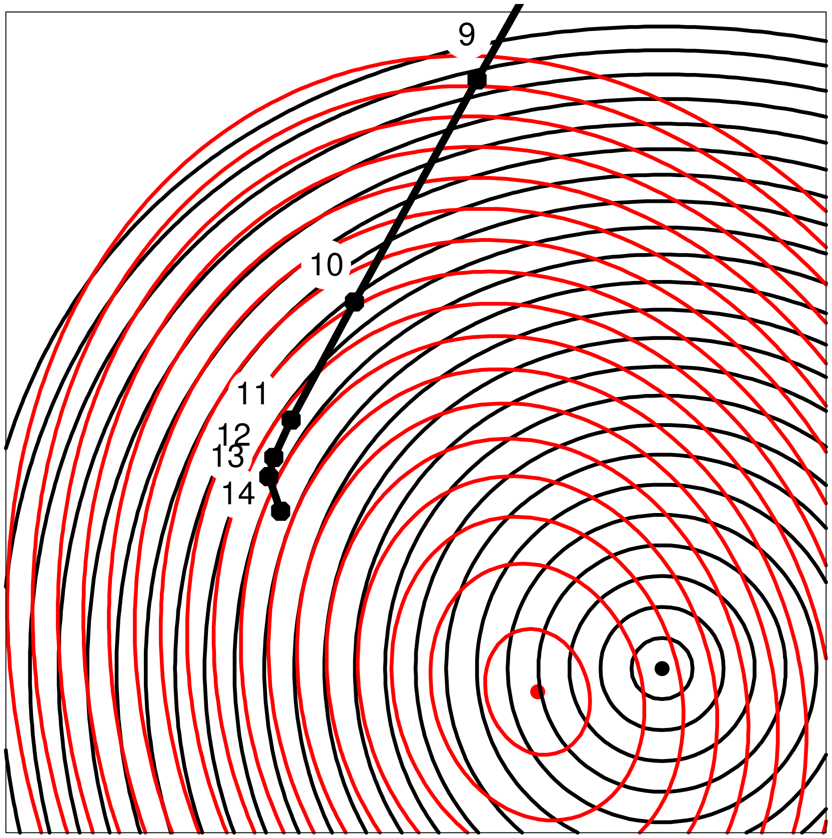
BFGS $(\lambda = 1,\ \hat{s} = 1)$ **SQN** $(\lambda = 0^+, \hat{s} \leq 1)$



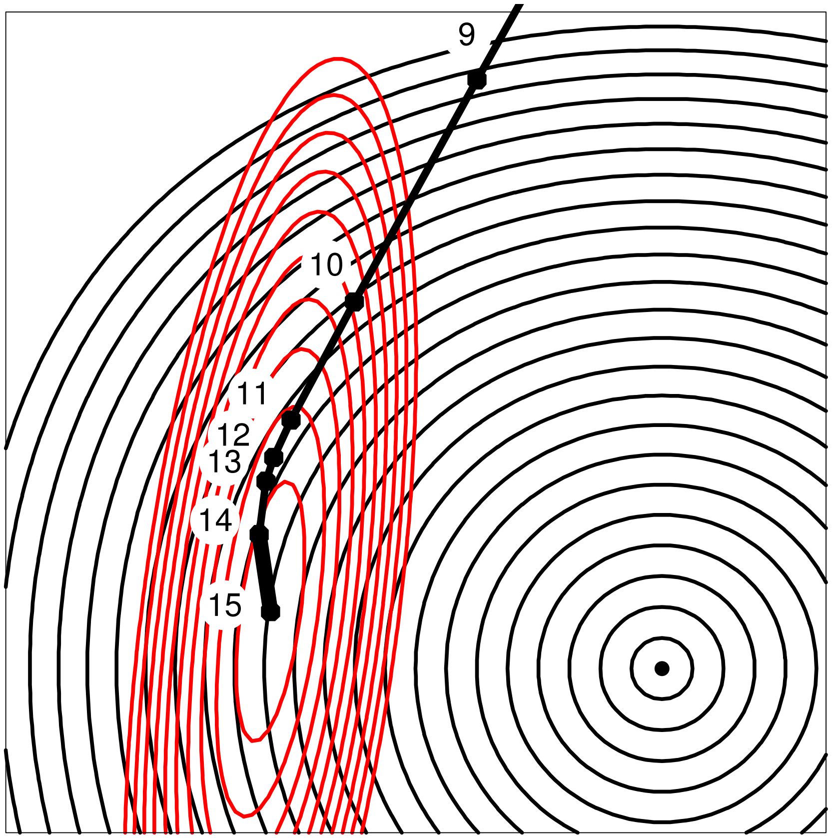
BFGS $(\lambda = 1,\ \hat{s} = 1)$



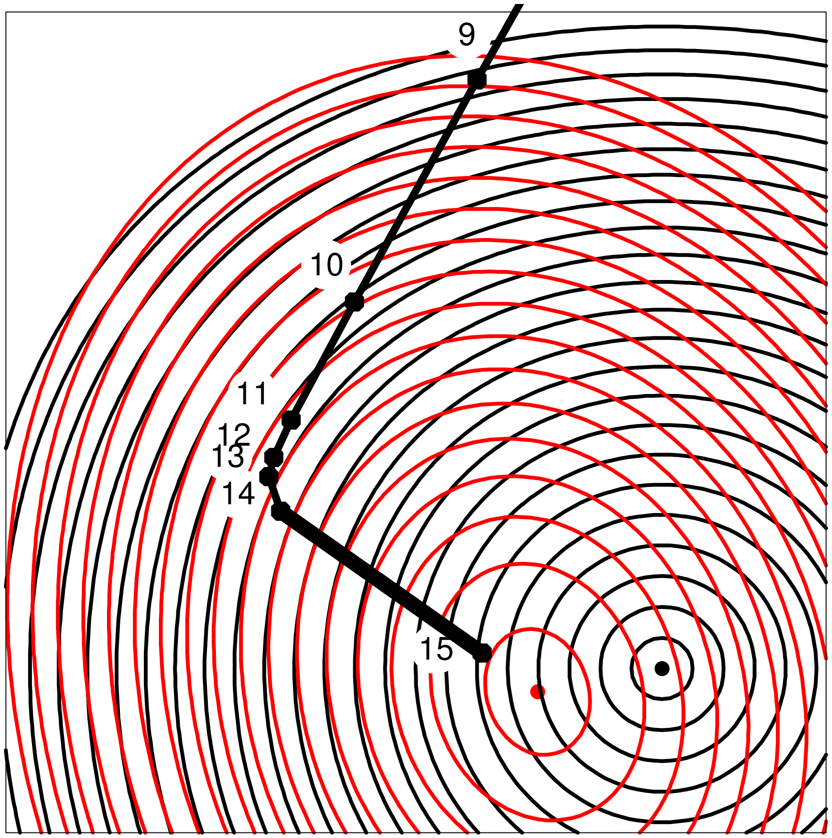
SQN $(\lambda = 0^+, \hat{s} \leq 1)$



BFGS $(\lambda = 1,\ \hat{s} = 1)$

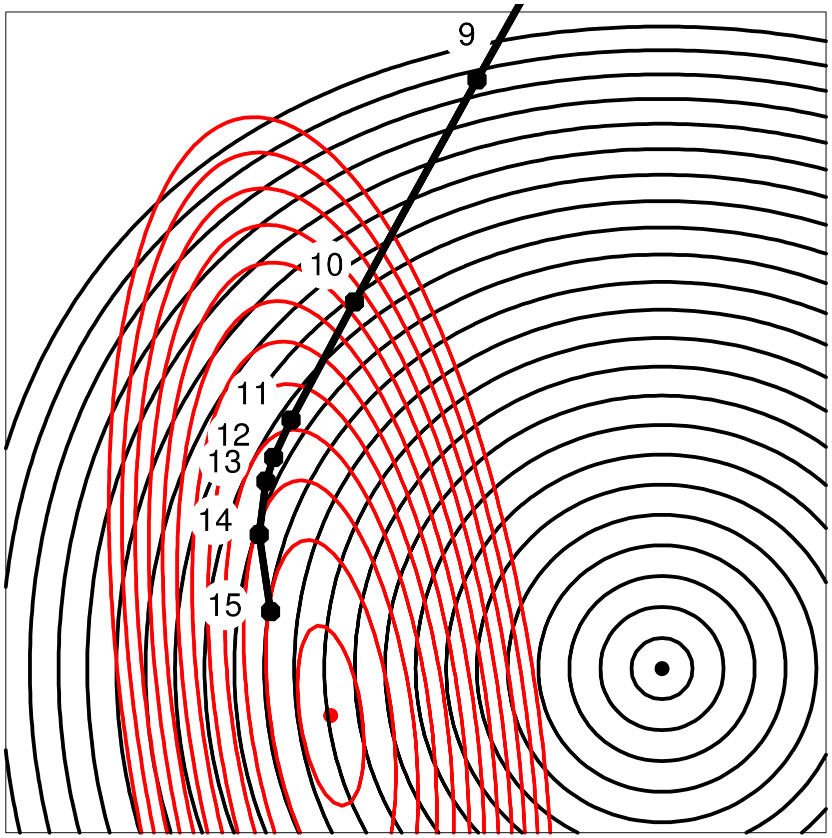


SQN $(\lambda = 0^+, \hat{s} \leq 1)$

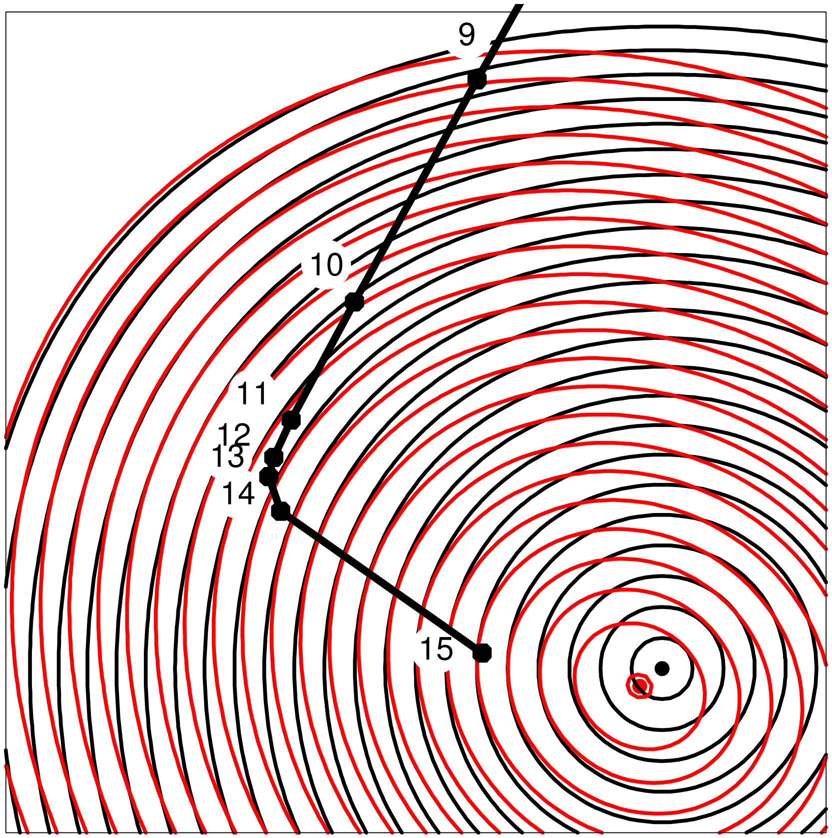


Minimize → Estimate Iteration 15

BFGS $(\lambda = 1,\ \hat{s} = 1)$

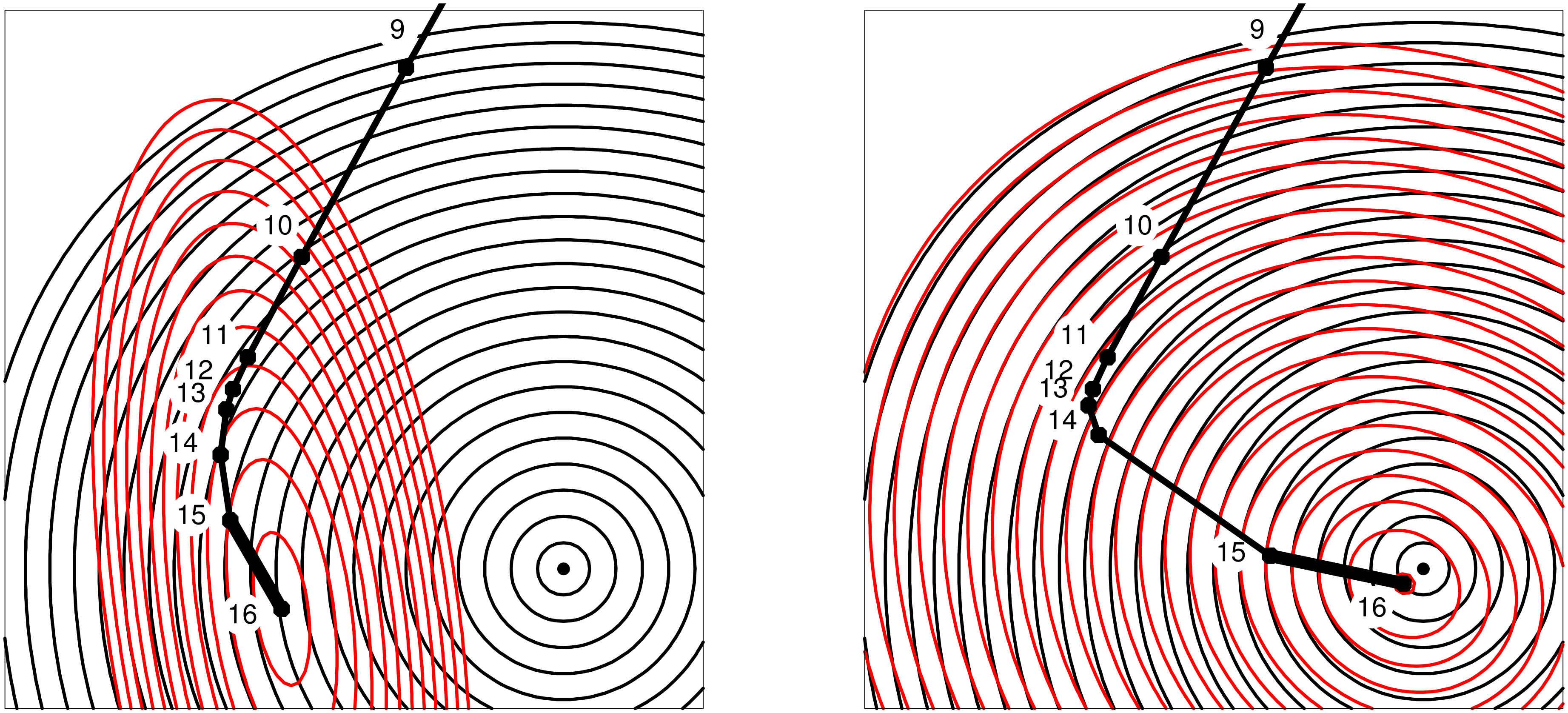


SQN $(\lambda = 0^+, \hat{s} \leq 1)$



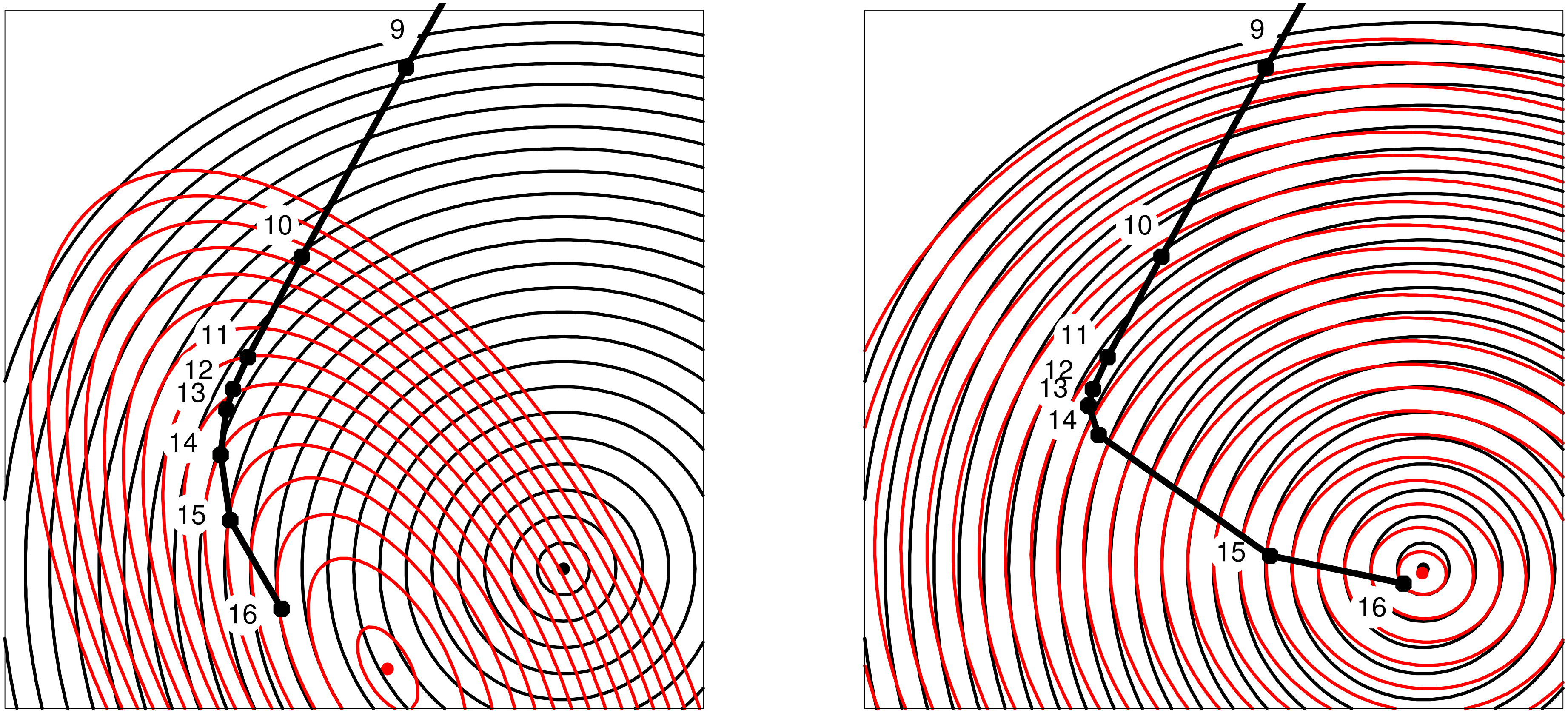
Minimize $\longrightarrow$ Estimate   Iteration 16

BFGS $(\lambda = 1,\ \hat{s} = 1)$ **SQN** $(\lambda = 0^{+}, \hat{s} \leq 1)$

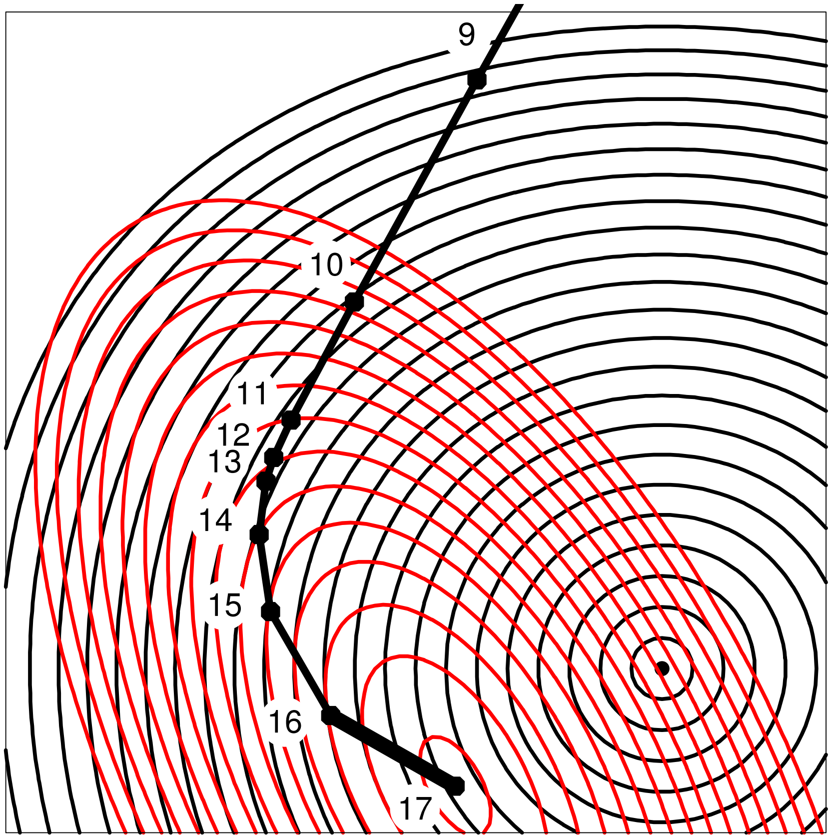


Minimize $\longrightarrow$ Estimate

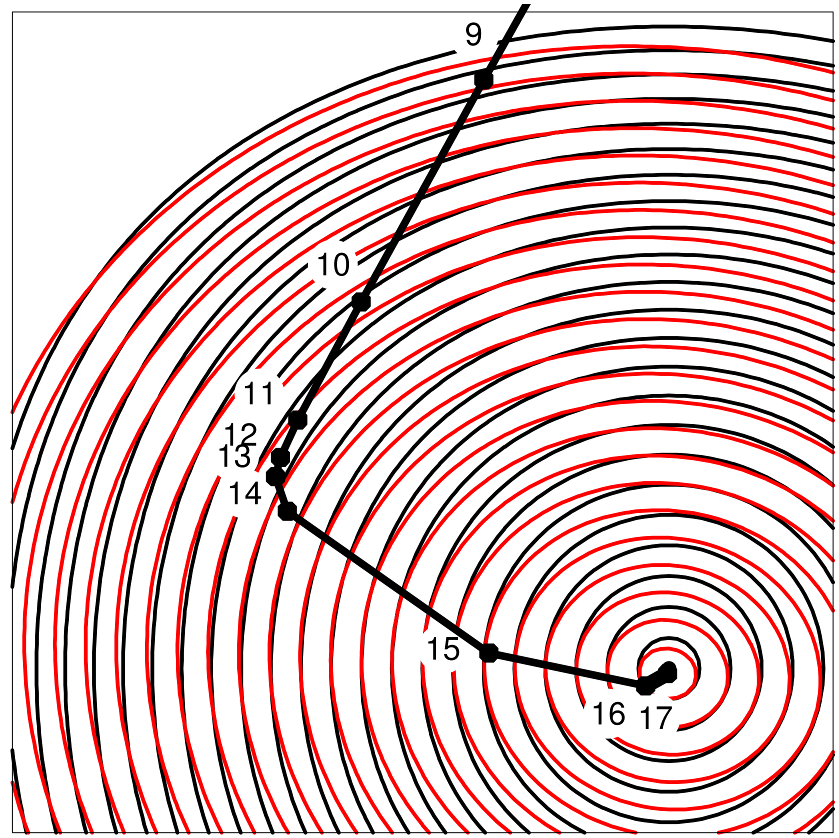
BFGS $(\lambda = 1,\ \hat{s} = 1)$ **SQN** $(\lambda = 0^+, \hat{s} \leq 1)$



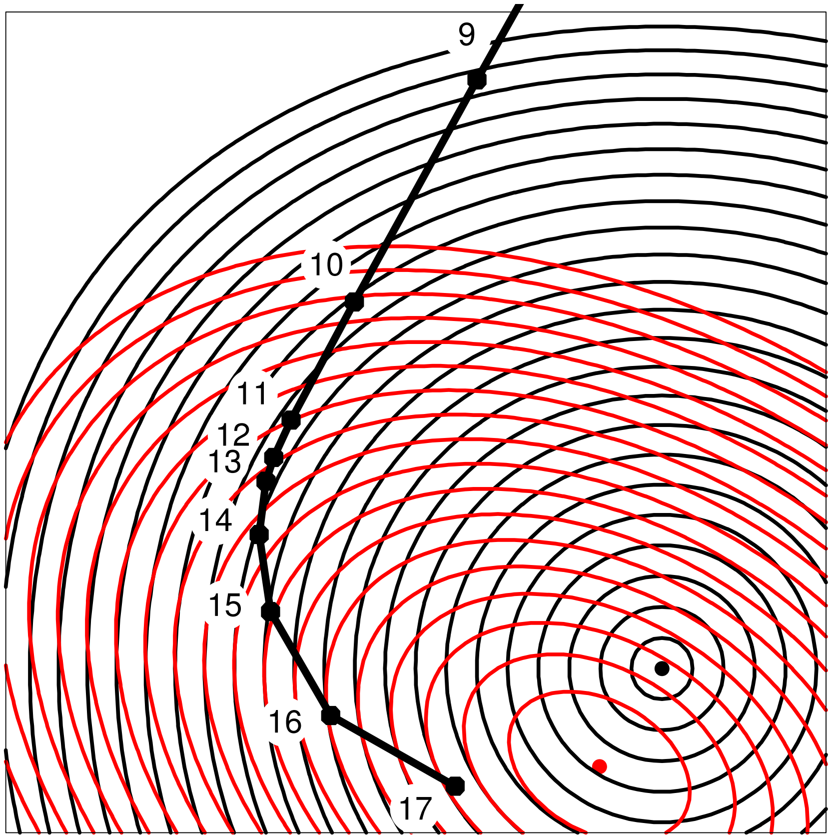
BFGS $(\lambda = 1,\ \hat{s} = 1)$



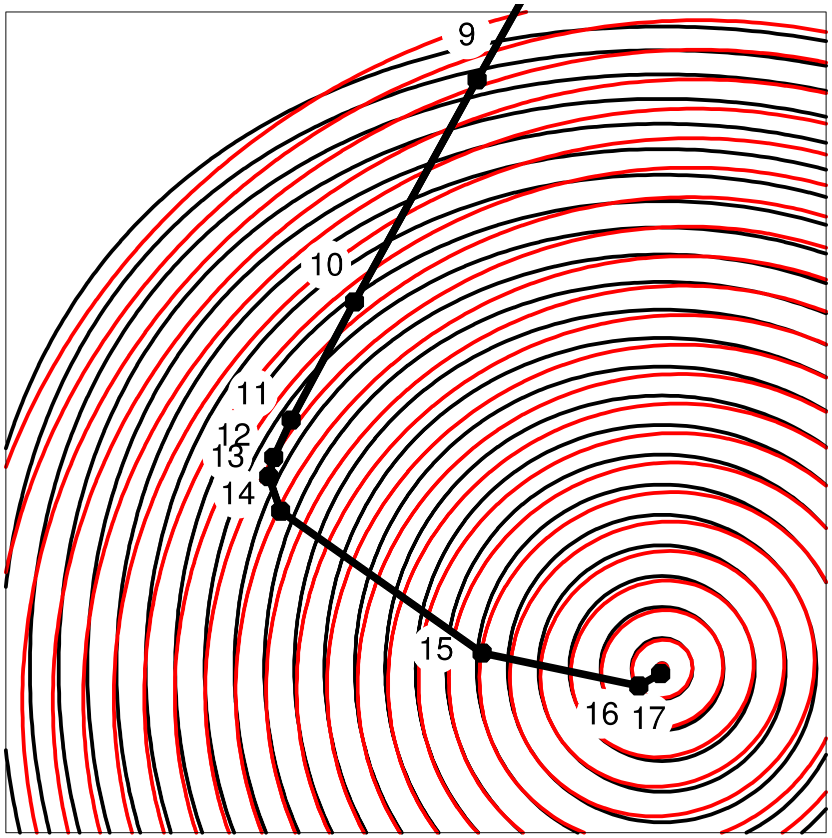
SQN $(\lambda = 0^+, \hat{s} \leq 1)$



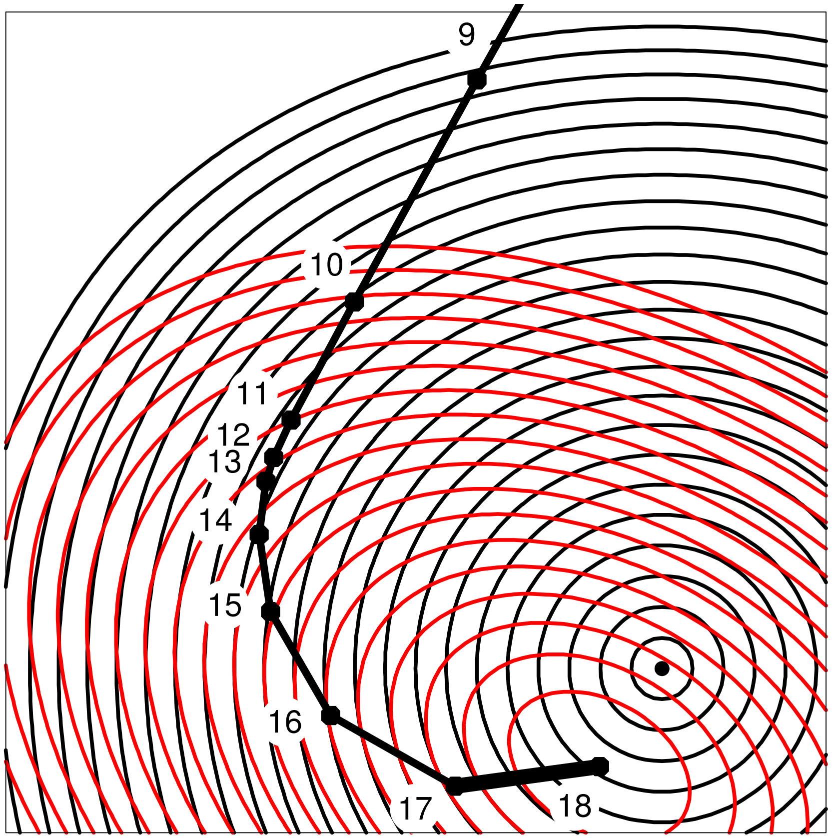
BFGS $(\lambda = 1,\ \hat{s} = 1)$



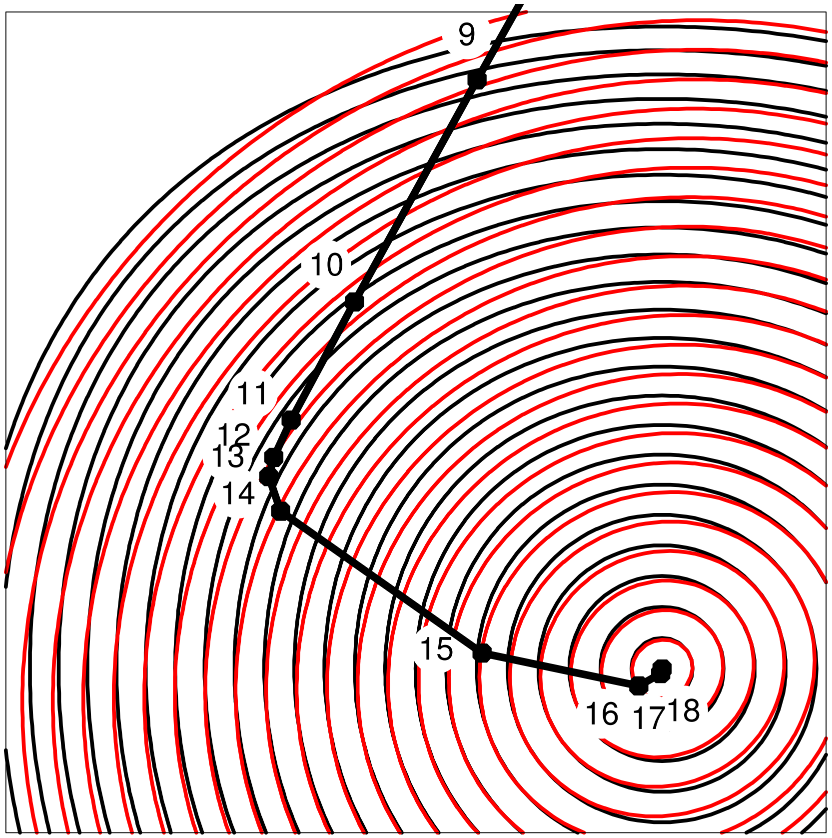
SQN $(\lambda = 0^+, \hat{s} \leq 1)$



BFGS $(\lambda = 1,\ \hat{s} = 1)$

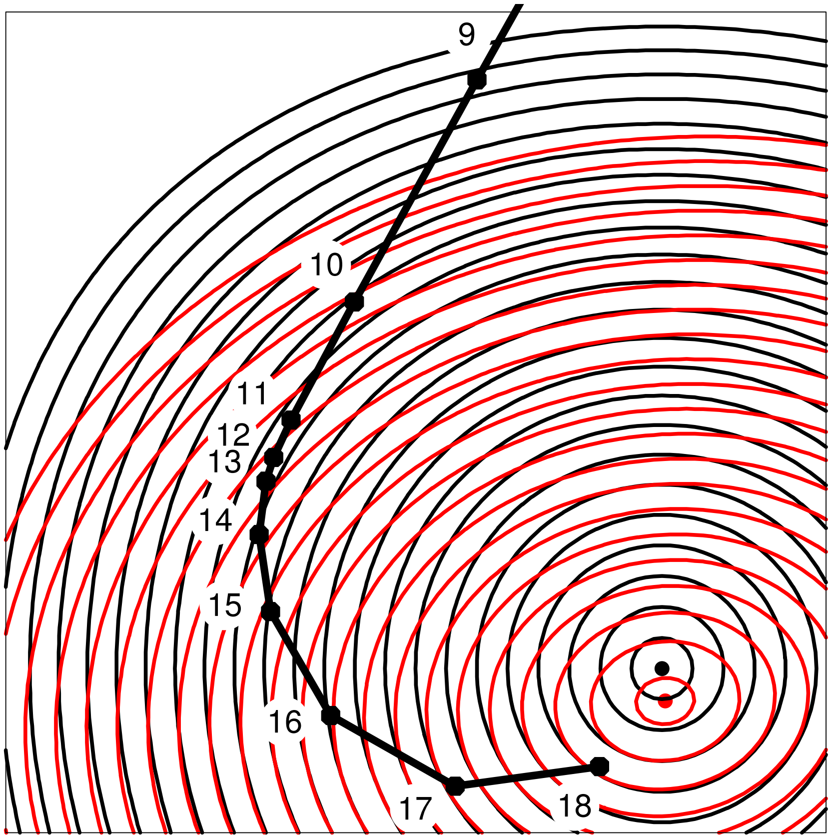


SQN $(\lambda = 0^+, \hat{s} \leq 1)$

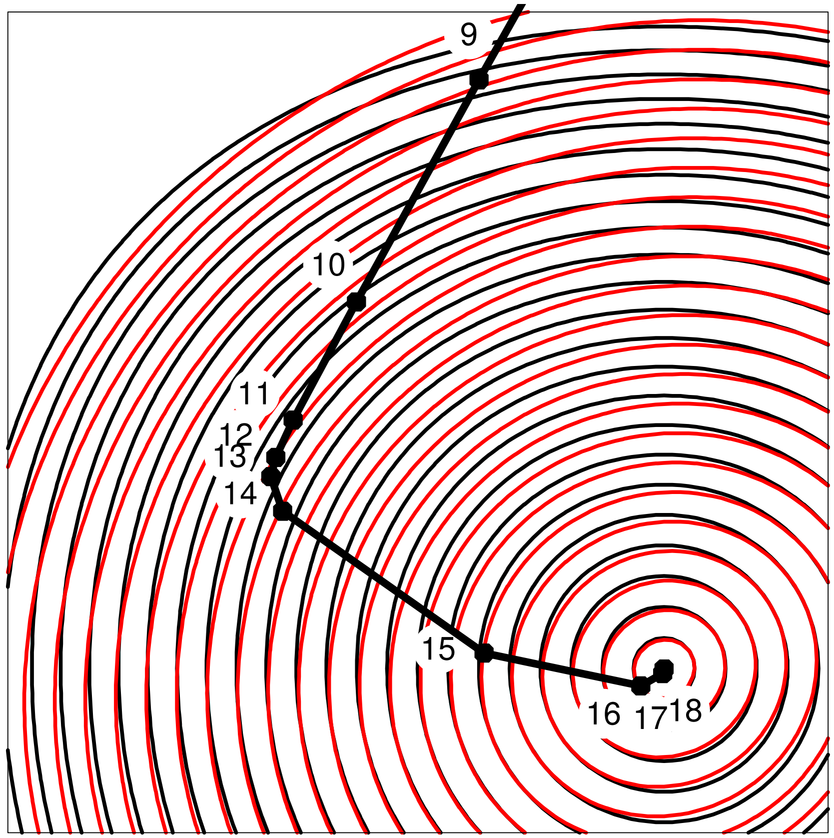


Minimize $\to$ Estimate Iteration 18

BFGS $(\lambda = 1,\ \hat{s} = 1)$

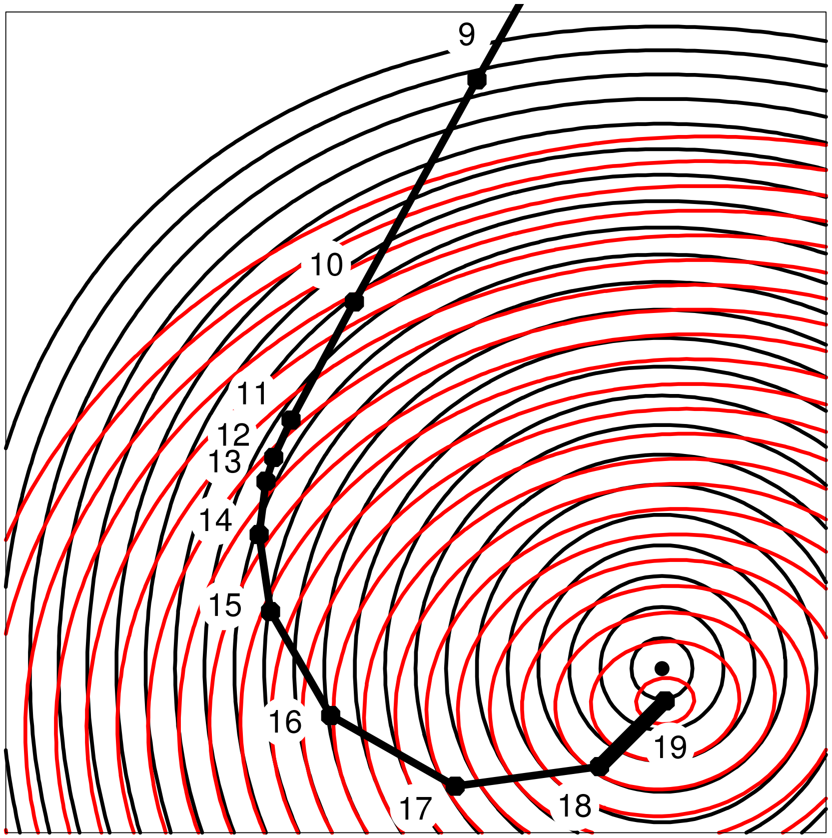


SQN $(\lambda = 0^+, \hat{s} \leq 1)$

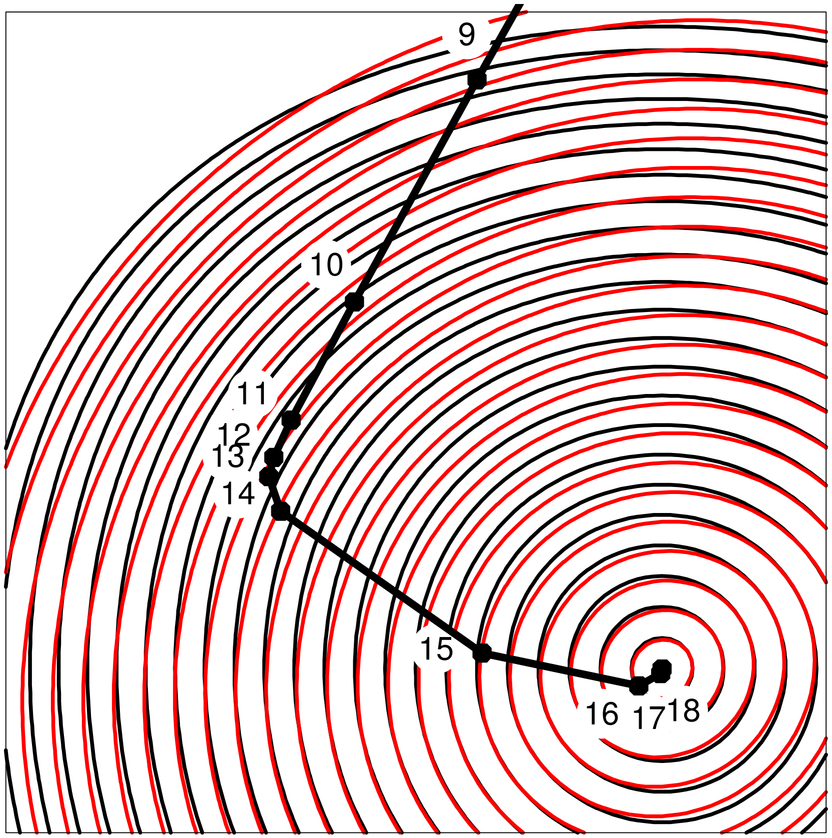


Minimize $\longrightarrow$ Estimate **Iteration 19**

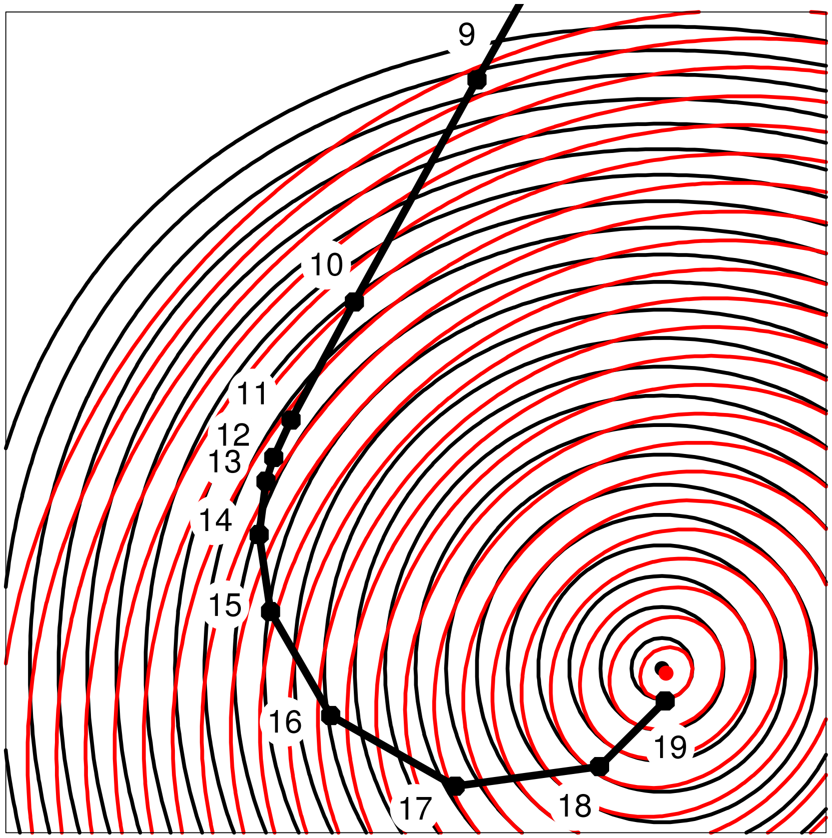
BFGS $(\lambda = 1,\ \hat{s} = 1)$



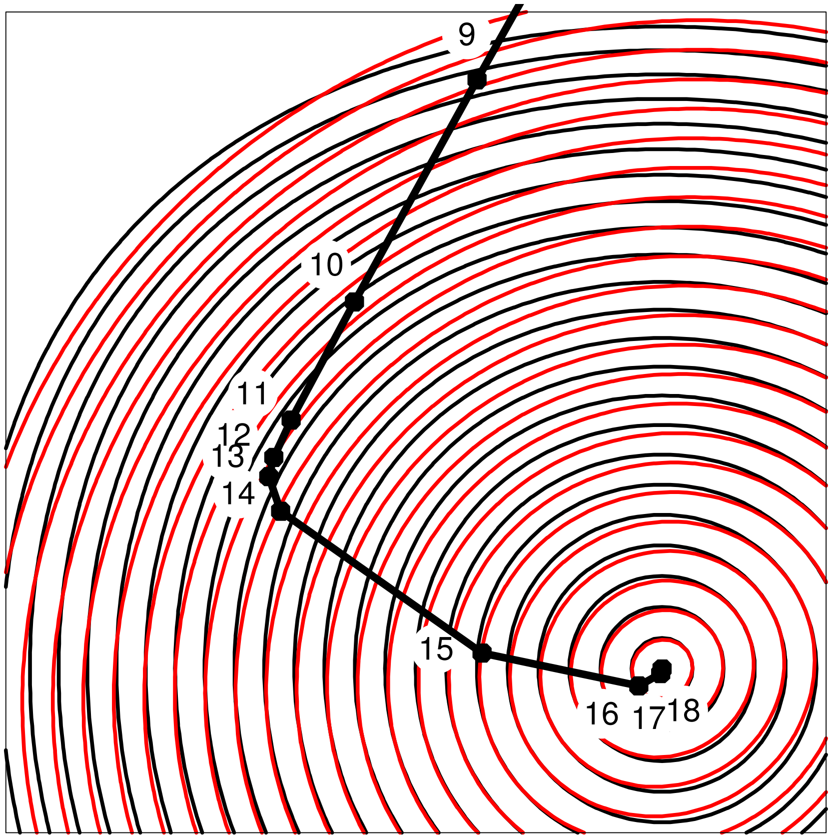
SQN $(\lambda = 0^+,\ \hat{s} \leq 1)$



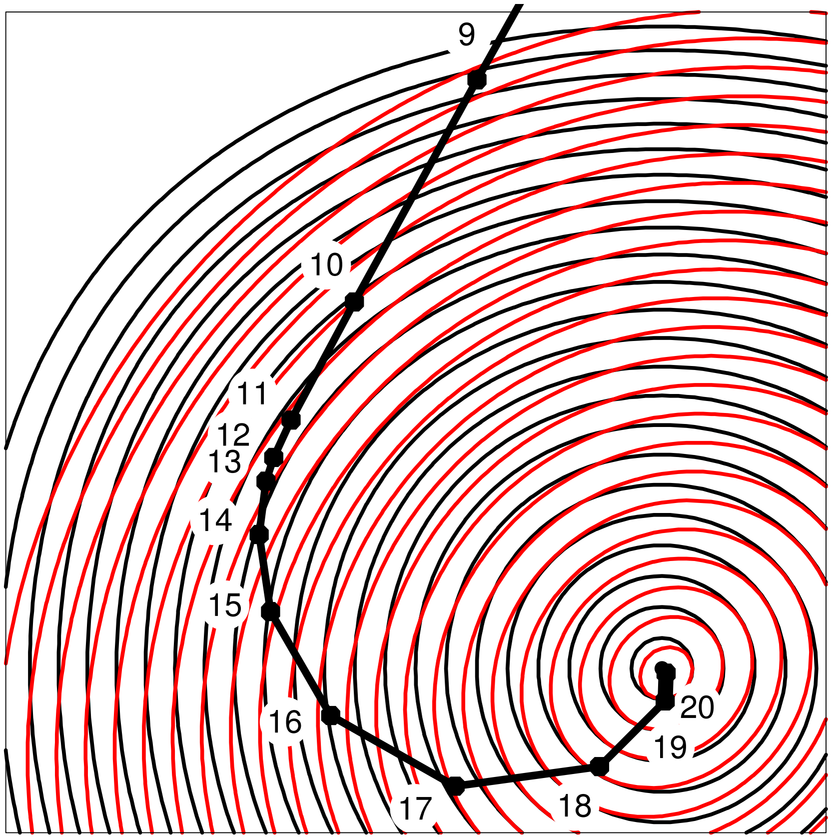
BFGS $(\lambda = 1,\ \hat{s} = 1)$



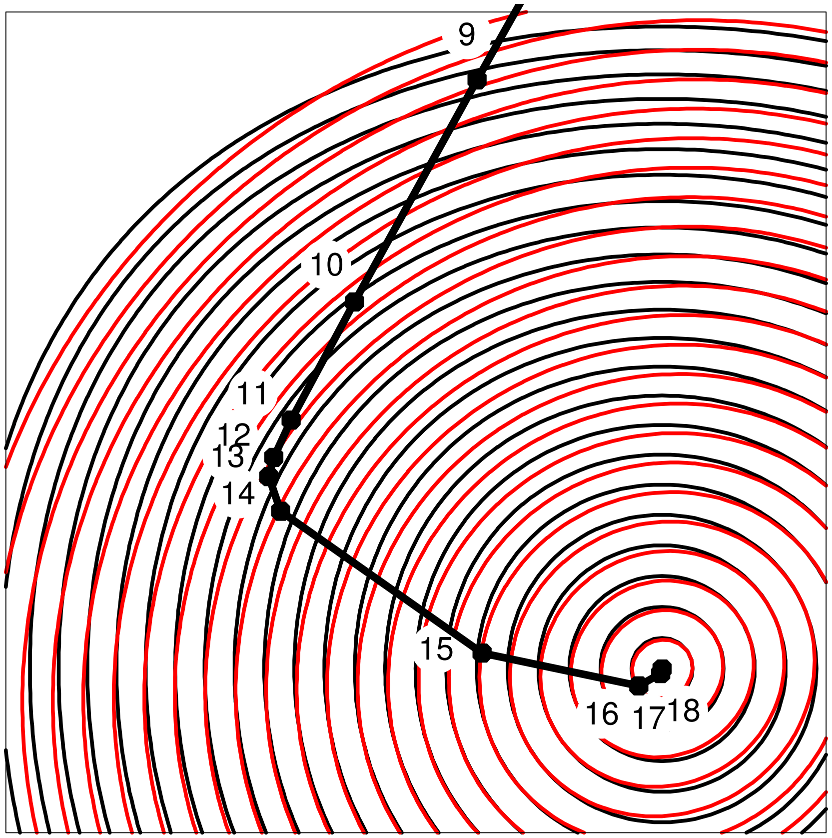
SQN $(\lambda = 0^+, \hat{s} \leq 1)$



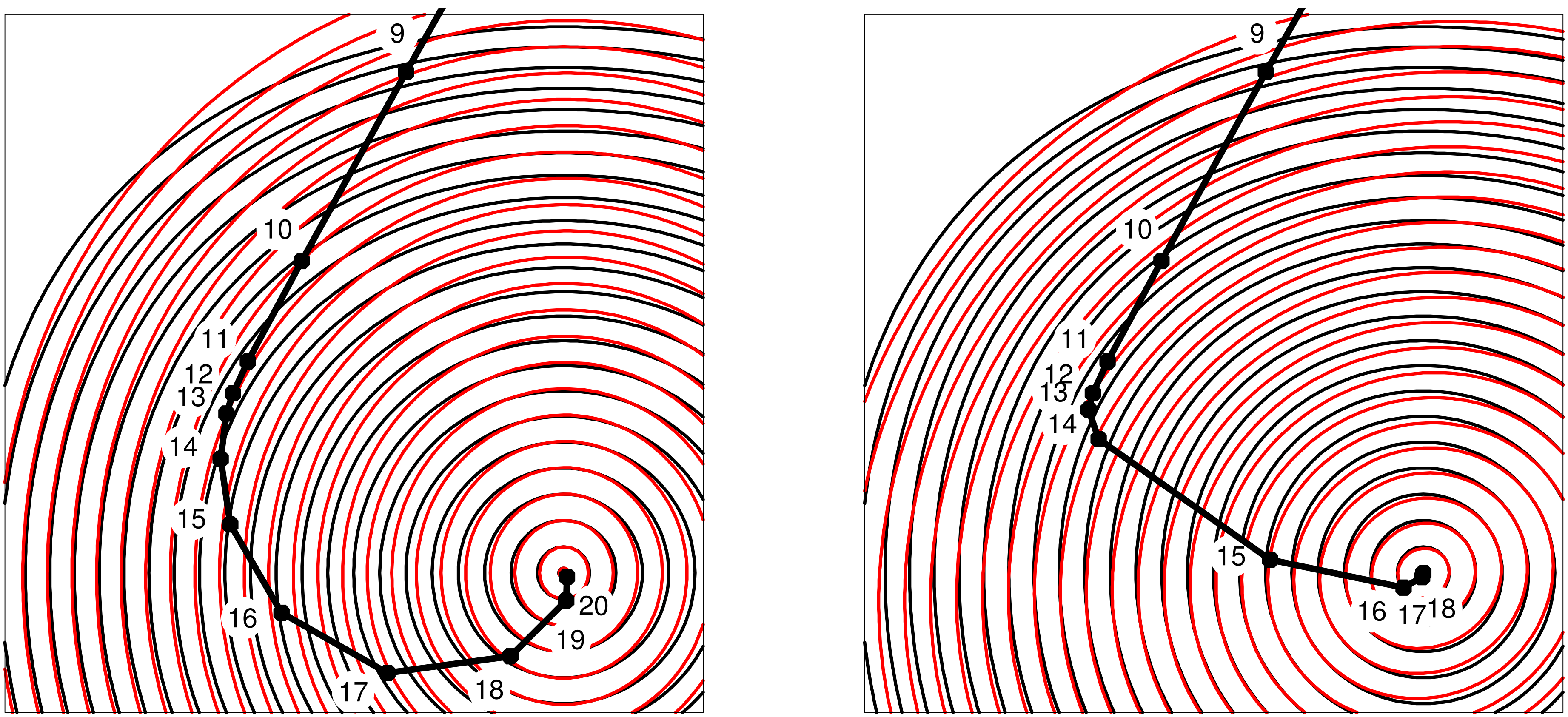
BFGS $(\lambda = 1,\ \hat{s} = 1)$



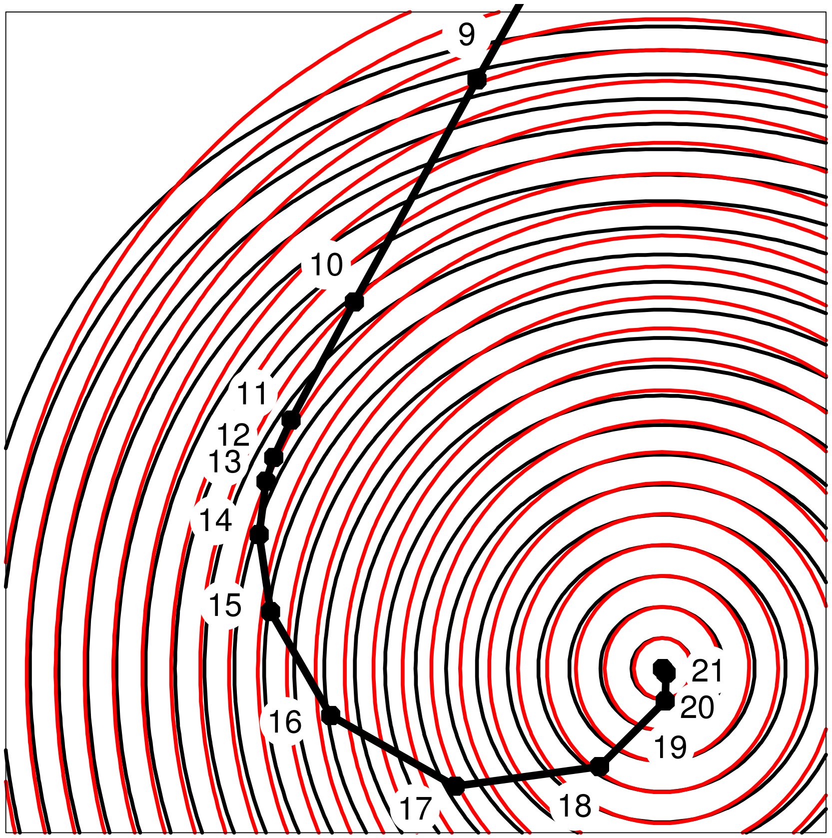
SQN $(\lambda = 0^+, \hat{s} \leq 1)$



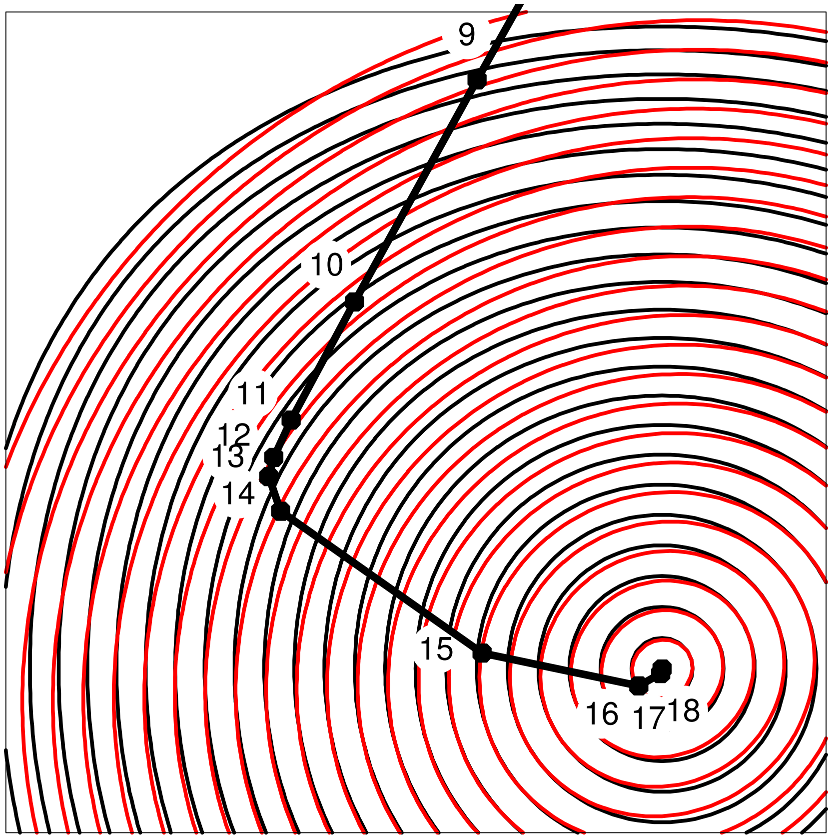
BFGS $(\lambda = 1, \hat{s} = 1)$ **SQN** $(\lambda = 0^+, \hat{s} \leq 1)$



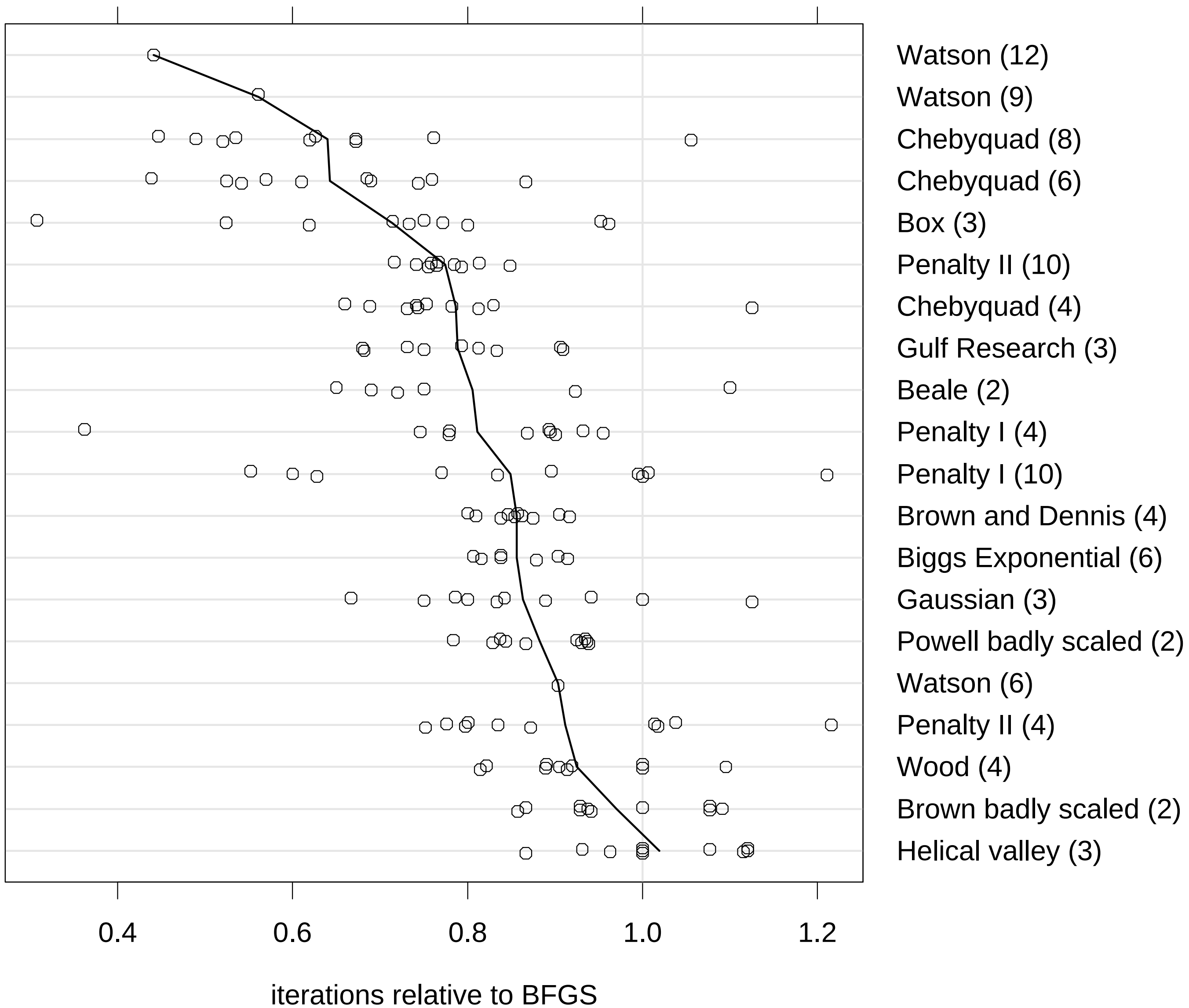
BFGS $(\lambda = 1,\ \hat{s} = 1)$



SQN $(\lambda = 0^+, \hat{s} \leq 1)$



Standard test problems: iteration counts



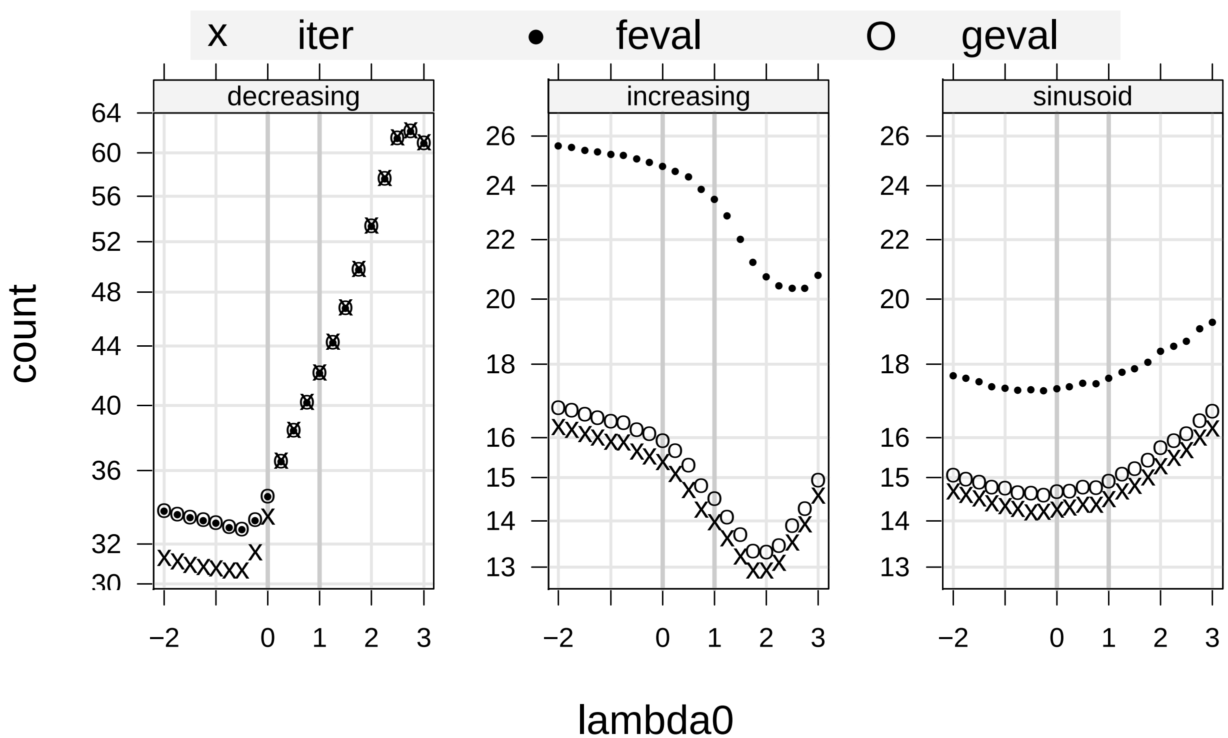
– Problem set: Moré, Garbow and Hillstrom, 1981

– Line search: Fletcher, 1987, pp.33-38

Three convex functions with diagonal Hessians.

Toward optimum, H	$H_{ii}(x)$	anticipated best λ
Decreases:	$1 + (\eta_i x_i)^2$	$\lambda < 0$
Increases:	$[1 + (\eta_i x_i)^2]^{-1}$	$\lambda > 0$
Sinusoid:	$[1 + \sin(\eta_i x_i)]$	$\lambda = 0$

$n = 4$, $\eta = (1, 2, 4, 8)$, and 1000 random starts



Hessian update

- New *Statistical Least-Change metric*
  - preserves accuracy in orthogonal complement, C
  - relies on future iterations to improve the complement
- Result is in the negative Broyden family

Steplength

- Wishart model captures uncertainty about the Hessian.
- Estimated step sizes work better

Statistical thinking $\Longrightarrow$ Improved performance over to BFGS