CS Space

Unconstrained Nonlinear Optimization
Lecture 7: Quasi-Newton Methods

Feodor Pisnitchenko

CMCC/UFABC

Momoe Sakamori

IMECC/UNICAMP

Summary

- Quasi-Newton Methods
- DFP
- BFGS
- SR1
- The Broyden Class
- Tests

Quasi-Newton Methods

- Consider the unconstrained optimization problem:

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

- and line search iteration:

$$x_{k+1} = x_k + \alpha_k p_k$$

  where
  - α_k: step length to be chosen according to the Wolfe conditions.
  - $p_k = -B_k^{-1} \nabla f_k$: search direction

- **Quasi-Newton methods**, like steepest descent, require only the **gradient** of the objective function at each iteration.

- They build a **model** of the objective function by **measuring changes in gradients**, achieving **superlinear convergence**.

Quasi-Newton Method

- Consider the quadratic approximation

$$m_k(p) = f_k + \nabla f_k^T p + \frac{1}{2} p^T B_k p$$

- $B_k \in \mathbb{R}^{n \times n}$ is symmetric positive definite and will be updated at each iteration.
- $\nabla_p m_k(p) = \nabla f_k + B_k p$
- Note that $m_k(0) = f_k$ and $\nabla_p m_k(0) = \nabla f_k$.
- The minimizer of $m_k(p)$ is the search direction

$$p_k = -B_k^{-1} \nabla f_k$$

Quasi-Newton Method

- Instead of **calculating** B_k from scratch at each iteration, Davidon proposed a simple **update** strategy.
- The update **incorporates curvature information** measured during the most recent step.
- After generating a new iterate $x_{k+1} = x_k + \alpha_k p_k$, the goal is to construct a new quadratic model based on the updated information:

$$m_{k+1}(p) = f_{k+1} + \nabla f_{k+1}^T p + \frac{1}{2} p^T B_{k+1} p.$$

The gradient is given by

$$\nabla m_{k+1}(p) = \nabla f_{k+1} + B_{k+1} p.$$

The secant equation

- What requirements should we impose on B_{k+1}?
- The gradient of m_{k+1} should match the gradient of f at x_k and x_{k+1}.
- At x_{k+1}: $\nabla m_{k+1}(0) = \nabla f_{k+1}$ is satisfied automatically.
- At $x_k = x_{k+1} - \alpha_k p_k$, we have

$$\begin{aligned} \nabla m_{k+1}(-\alpha_k p_k) &= \nabla f_{k+1} - \alpha_k B_{k+1} p_k = \nabla f_k \\ &\Rightarrow B_{k+1} \alpha_k p_k = \nabla f_{k+1} - \nabla f_k \end{aligned}$$

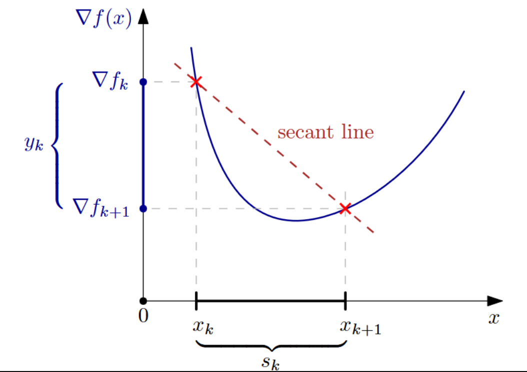
- Then we have the **secant equation**:

$$B_{k+1} s_k = y_k$$

where $s_{k+1} = x_{k+1} - x_k = \alpha_k p_k$ and $y_k = \nabla f_{k+1} - \nabla f_k$.

The secant equation

- Reason for this name: for a real-valued function of one real variable $g : \mathbb{R} \to \mathbb{R}$, the ratio y_k/s_k is the slope of a secant line joining two points on a curve.



Curvature condition

- Secant equation: $B_{k+1}s_k = y_k$.
- Such positive definite matrix B_{k+1} is possible only if the following curvature condition holds:

$$s_k^T y_k > 0.$$

- When f is strongly convex, the curvature condition will be satisfied for any two point x_k and x_{k+1}, but it is not always the case for nonconvex functions.
- The curvature condition is guaranteed to hold if the (strong) Wolfe conditions are imposed on the line search.

Choice of B_{k+1}

- When the curvature condition

$$y_k^T s_k \geq (c_2 - 1)\alpha_k \nabla f_k^T p_k$$

  is satisfied, the secant equation has infinitely many solutions.
- To determine B_{k+1} uniquely, among all symmetric matrices satisfying the secant equation, we impose an additional condition.

Choice of B_{k+1}

- We seek for B_{k+1} that is "closest" to B_k:

$$\begin{aligned} \min_{B \in \mathbb{R}^{n \times n}} \quad & \|B - B_k\| \\ \text{s.t.} \quad & B = B^T \\ & B s_k = y_k \end{aligned}$$

  where $s_k = x_{k+1} - x_k$ and $y_k = \nabla f_{k+1} - \nabla f_k$ and B_k is symmetric positive definite.

- Different matrix norms can be used, leading to a different quasi-Newton method.

Weighted Frobenius norm

- A useful norm is the weighted Frobenius norm

$$\|A\|_W = \|W^{1/2} A W^{1/2}\|_F$$

  where $\|.\|_F$ is defined by $\|C\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}^2$.
- W is any matrix that satisfies $W y_k = s_k$.
- Specifically, let $W = G_k^{-1}$ where G_k is the average Hessian defined by

$$G_k = \int_0^1 \nabla^2 f(x_k + \tau \alpha_k p_k) d\tau$$

- From Taylor's theorem, we ended up with

$$y_k = G_k \alpha_k p_k = G_k s_k$$

DFP updating formula for B_k

- Using the weighted Frobenius norm with the average inverse Hessian, the unique solution is

$$\text{(DFP)} \qquad B_{k+1} = \left(I - \frac{y_k s_k^T}{y_k^T s_k}\right) B_k \left(I - \frac{s_k y_k^T}{y_k^T s_k}\right) + \frac{y_k y_k^T}{y_k^T s_k}.$$

- It was originally proposed by **D**avidon, and subsequently studied by **F**letcher and **P**owell.

Sherman-Morrison-Woodbury Formula

- Let $A \in \mathbb{R}^{n\times n}$ and $U, V \in \mathbb{R}^{p\times n}$, with $1 \le p \le n$.
- Define $\hat{A} = A + UV^T$, then $\hat{A}$ is nonsingular if and only if $(I + V^T A^{-1} U)$ is nonsingular.
- In this case,

$$\hat{A}^{-1} = A^{-1} - A^{-1}U(I + V^T A^{-1} U)^{-1} V^T A^{-1}.$$

DFP updating formula for H_k

- Let $B_k^{-1} = H_k$.
- Using the Sherman-Morrison-Woodbury formula, we have the update of H_k that corresponds to the DFP update of B_k:

$$\text{(DFP)} \qquad H_{k+1} = H_k - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k} + \frac{s_k s_k^T}{y_k^T s_k}$$

- Both H_k and B_k undergo rank-two updates.
- Quasi-Newton updating aims to update Hessian (or inverse Hessian) approximations without full recomputation.

BFGS Method

- Named after its developers: **B**royden, **F**letcher, **G**oldfarb, and **S**hanno.
- BFGS is derived by altering the approach used by DFP: rather than imposing conditions on Hessians B_k, it focuses on their inverses H_k.
- In BFGS, H_k is directly approximated by solving

$$
\begin{aligned}
\min_{H \in \mathbb{R}^{n\times n}} \quad & \|H - H_k\| \\
\text{s.t.} \quad & H = H^T \\
& H y_k = s_k
\end{aligned}
$$

- The norm is again weighted Frobenius with $W s_k = y_k$.

BFGS updating matrices

- Let $W = G_k$, the update inverse approximation is given by

$$\text{(BFGS)} \qquad H_{k+1} = \left(I - \frac{s_k y_k^T}{y_k^T s_k}\right)^T H_k \left(I - \frac{y_k s_k^T}{y_k^T s_k}\right) + \frac{s_k s_k^T}{y_k^T s_k}.$$

- If H_k is positive definite, then H_{k+1} is positive definite.
- The update formula for B_k is obtained by applying the Sherman-Morrison-Woodbury formula,

$$\text{(BFGS)} \qquad B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}.$$

How to choose H_0?

- Unfortunately, there is no universal choice that works well in all cases.
- H_0 is commonly set to be some multiple of the identity matrix.
- Another usual choice is to use a scaled identity matrix, where the scaling reflects the variable magnitudes.
- A useful trick is to scale the initial matrix after the first step is taken, but before applying the first BFGS update:

$$H_0 \leftarrow \frac{y_k^T s_k}{y_k^T y_k} I.$$

Algorithm BFGS Method

- The cost of each iteration: $\mathcal{O}(n^2)$ arithmetic operations (excluding function and gradient evaluations).
- The algorithm avoids $\mathcal{O}(n^3)$ operations like linear system solves and matrix–matrix multiplications.
- BFGS is robust and has a superlinear convergence rate, which is sufficient for most applications.
- Newton's method has faster (quadratic) convergence, but its cost per iteration usually is higher (its need for second derivatives and solution of a linear system).

Global Convergence of the BFGS Method

Theorem (Global Convergence)

Suppose that $f \in C^2$. Let B_0 be any symmetric positive definite matrix, and let x_0 be a starting point for which the sublevel set $L = \{x | f(x) \leq f(x_0)\}$ is convex. Furthermore, there exists $0 < m \leq M$ such that $m||z||^2 \leq z^T \Gamma(x) z \leq M||z||^2$ for all $z \in \mathbb{R}^n$ and $x \in L$ with $\Gamma(x) = \nabla^2 f(x)$. Then the sequence $\{x_k\}$ generated by the BFGS Method (with $\epsilon = 0$) converges to the minimizer x^* of f.

Superlinear Convergence of the BFGS Method

Theorem (Superlinear Convergence)

Suppose that $f \in C^2$ and that the iterates generated by the BFGS algorithm converge to a minimizer x^* at which $\Gamma(x)$ is locally Lipschitz. Suppose also that

$$\sum_{k=1}^{\infty} \|x_k - x^*\| < \infty.$$

Then x_k converges to x^* at a superlinear rate.

The SR1 Method

- In BFGS and DFP, the updated matrix B_{k+1} (or H_{k+1}) differs from B_k (or H_k) by a rank-2 matrix.
- There exists a rank-1 update, known as the Symmetric Rank-1 (SR1) update.
- The SR1 update maintains symmetry and satisfies the secant equation.
- Unlike BFGS and DFP, SR1 does not ensure positive definiteness of the updated matrix.
- Despite this, SR1-based algorithms have shown good numerical performance.

The SR1 Method

- The only symmetric rank-1 updating formula that satisfies the secant equation is given by

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)(y_k - B_k s_k)^T}{(y_k - B_k s_k)^T s_k}$$

- By applying the Sherman-Morrison formula, the corresponding update formula for the inverse Hessian approximation H_k is given by

$$H_{k+1} = H_k + \frac{(s_k - H_k y_k)(s_k - H_k y_k)^T}{(s_k - H_k y_k)^T y_k}$$

Convergence to the true Hessian

Theorem

Suppose that $f \in C^2$, and that its Hessian is bounded and Lipschitz continuous in a neighborhood of a point x^*. Let $\{x_k\}$ be any sequence of iterates such that $x_k \to x^*$ for some $x^* \in \mathbb{R}^n$. Suppose in addition that the inequality $|s_k^T(y_k - B_k s_k)| \geq r\|s_k\|\|y_k - B_k s_k\|$ holds for all k, for some $r \in (0,1)$, and that the steps s_k are uniformly linearly independent. Then the matrices B_k generated by the SR1 updating formula satisfy

$$\lim_{k\to\infty} \|B_k - \nabla^2 f(x^*)\| = 0.$$

The Broyden Class

- The **Broyden class** is a family of updates specified by the following general formula:

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k} + \phi_k (s_k^T B_k s_k) v_k v_k^T$$

where ϕ_k is a scalar parameter, and

$$v_k = \left[\frac{y_k}{y_k^T s_k} - \frac{B_k s_k}{s_k^T B_k s_k} \right].$$

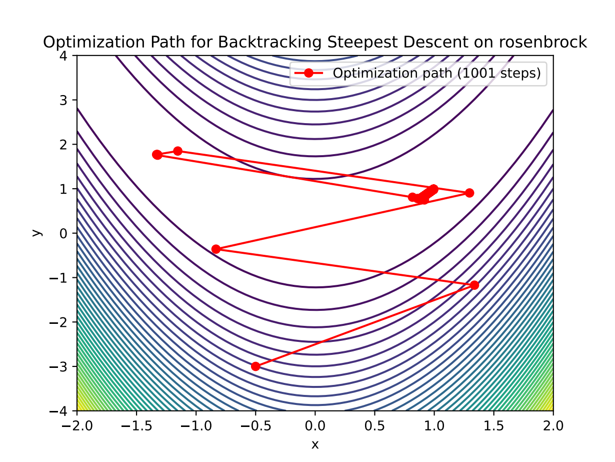
- BFGS: $\phi_k = 0$
- DFP: $\phi_k = 1$
- SR1: $\phi_k = \frac{s_k^T y_k}{s_k^T y_k - s_k^T B_k s_k}$ (It is a member of the Broyden class, but it does not belong to the restricted Broyden class.)

The Broyden Class

- Let $f : \mathbb{R}^n \to \mathbb{R}$ be $f(x) = \frac{1}{2} x^T A x + b^T x$, $A \succ 0$, and suppose B_0 is any symmetric positive definite matrix.
- Suppose α_k is the exact minimizer of $\phi_k(\alpha) = f(x_k + \alpha p_k)$, where $p_k = -B_k^{-1} \nabla f(x_k)$.
- B_k is updated by DFP or BFGS formula, and assume $s_k^T y_k \neq 0$ for all k (no update is skipped).
- In at most n iterations, the iterates converge to the solution.
- If n iterations are performed, we have $B_n = A$.
- For SR1, we obtain similar result, but assuming that $(s_k - H_k y_k)^T y_k \neq 0$ for all k.

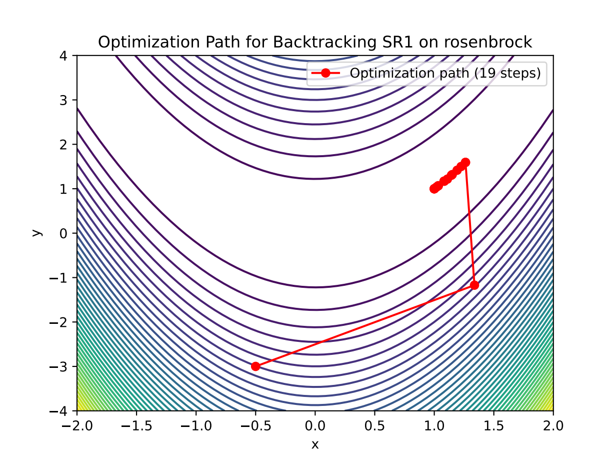
Example 1

$$F(x, y) = (1 - x)^2 + 100(y - x^2)^2$$



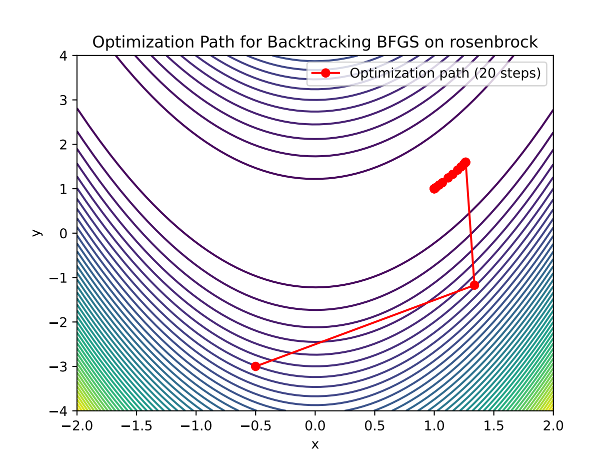
Example 1

$$F(x,y) = (1-x)^2 + 100(y-x^2)^2$$



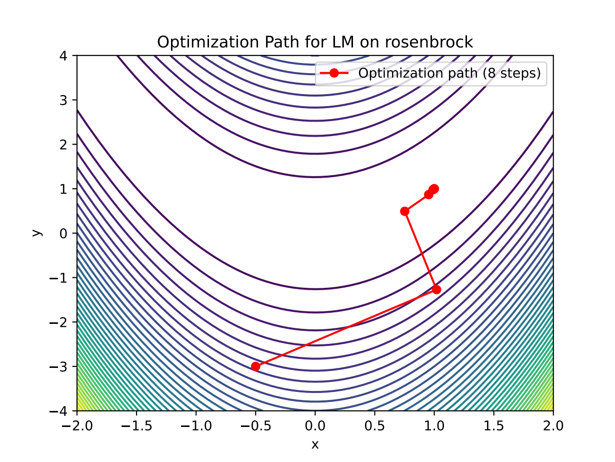
Example 1

$$F(x, y) = (1 - x)^2 + 100(y - x^2)^2$$



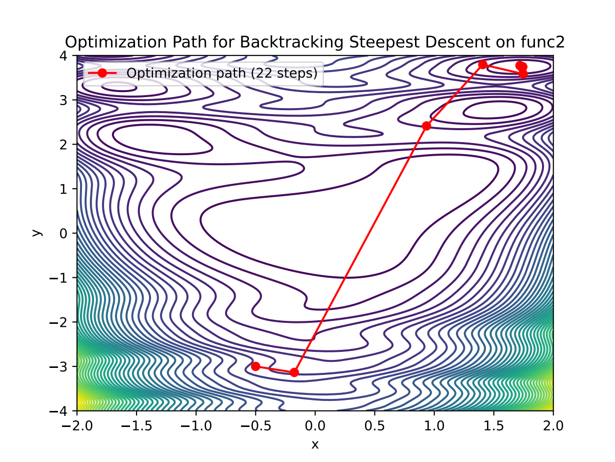
Example 1

$$F(x,y) = (1-x)^2 + 100(y-x^2)^2$$



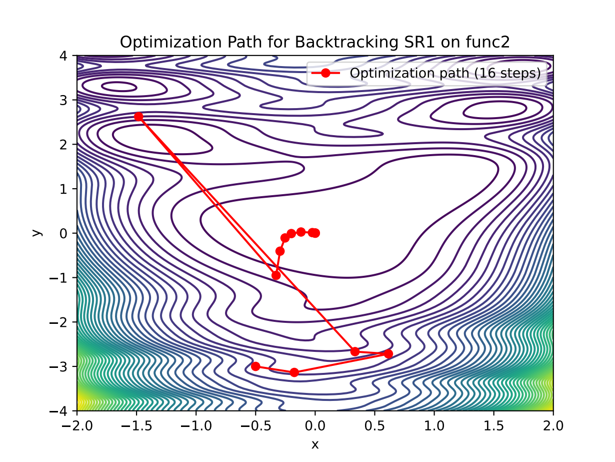
Example 2

$$F(x,y) = (1 - x^4)^2 + 100(y - x^2)^2 + 100(\sin(y^2) - x)^2$$



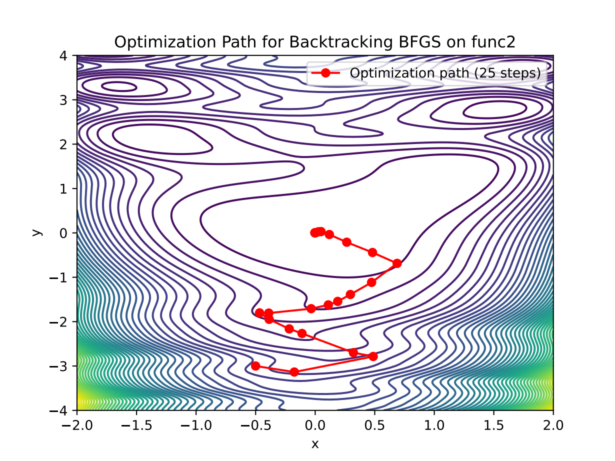
Example 2

$$F(x,y) = (1 - x^4)^2 + 100(y - x^2)^2 + 100(\sin(y^2) - x)^2$$



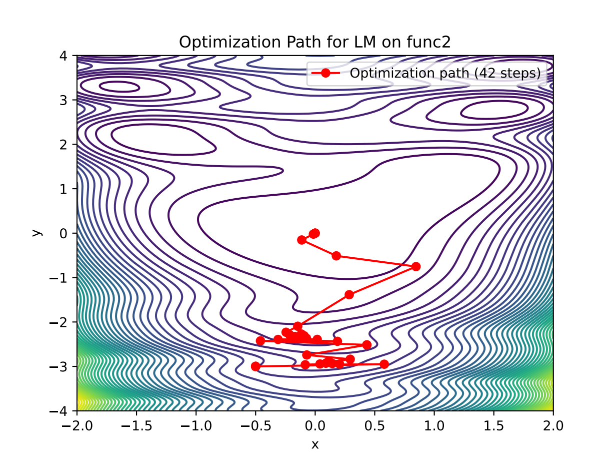
Example 2

$$F(x,y) = (1 - x^4)^2 + 100(y - x^2)^2 + 100(\sin(y^2) - x)^2$$



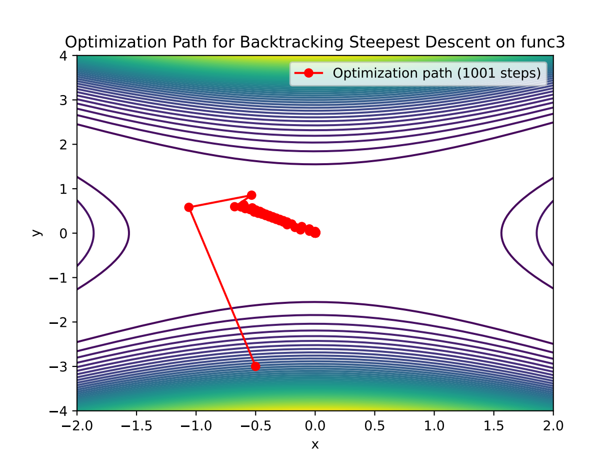
Example 2

$$F(x,y) = (1 - x^4)^2 + 100(y - x^2)^2 + 100(\sin(y^2) - x)^2$$



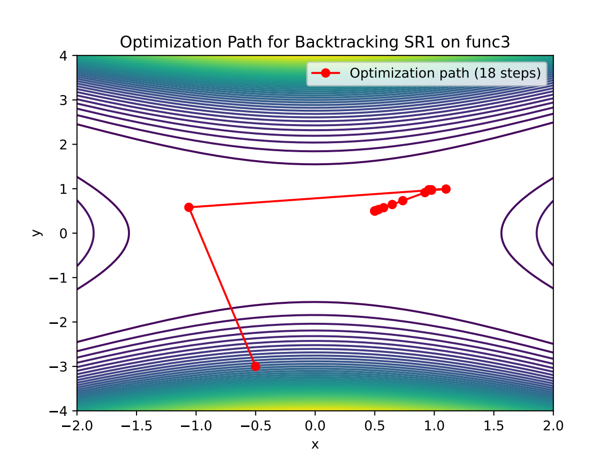
Example 3

$$F(x, y) = (2y^2 - x)^2 + 100(y^2 - x^2)^2$$



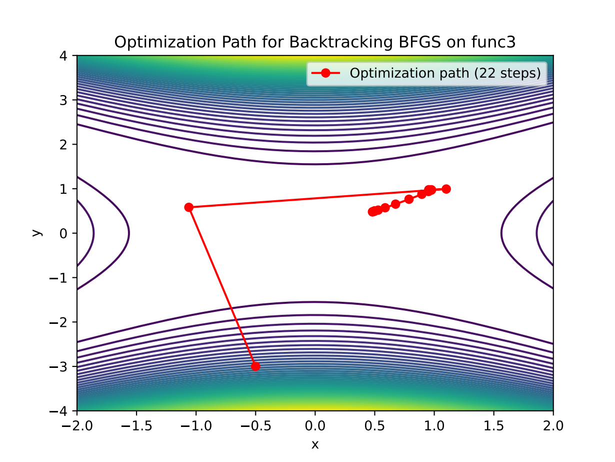
Example 3

$$F(x, y) = (2y^2 - x)^2 + 100(y^2 - x^2)^2$$



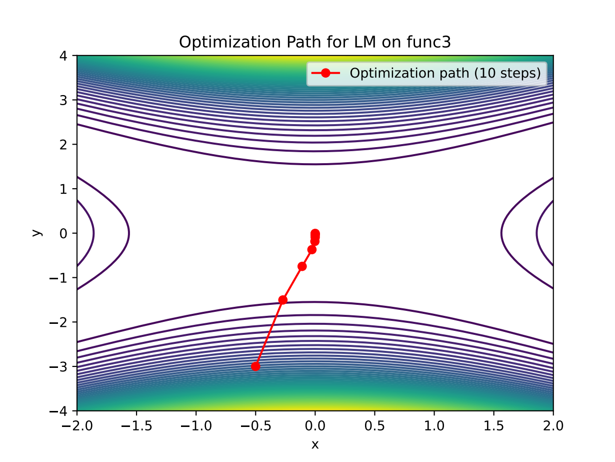
Example 3

$$F(x, y) = (2y^2 - x)^2 + 100(y^2 - x^2)^2$$



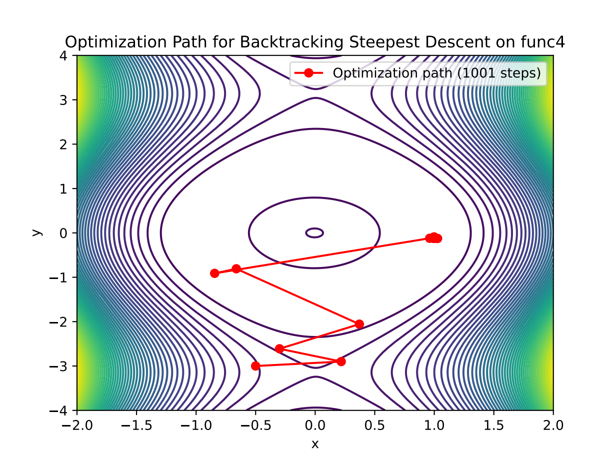
Example 3

$$F(x, y) = (2y^2 - x)^2 + 100(y^2 - x^2)^2$$



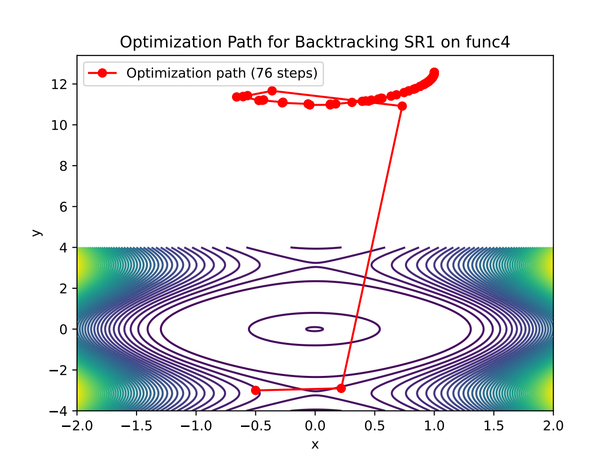
Example 4

$$F(x, y) = (1 - \sin(x))^2 + 100(\cos(y) - x^2)^2$$



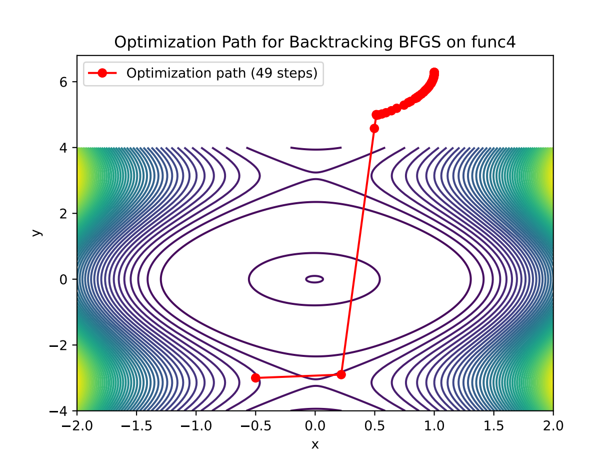
Example 4

$$F(x, y) = (1 - \sin(x))^2 + 100(\cos(y) - x^2)^2$$



Example 4

$$F(x, y) = (1 - \sin(x))^2 + 100(\cos(y) - x^2)^2$$



Example 4

$$F(x, y) = (1 - \sin(x))^2 + 100(\cos(y) - x^2)^2$$