Optimization Algorithms

Simple cases

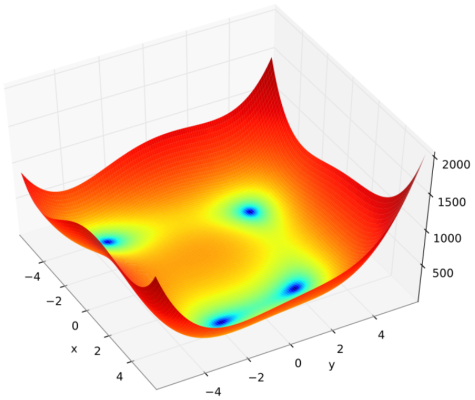
- Local minima
- Reasonable initial guess
- Wide attractive basins

Hard cases

- Noisy function evaluations
- High dimensionality

Popular representatives

- Newton
- Steepest descent
- BFGS
- L-BFGS



The Himmelblau function

Newton's method

$$\underbrace{\mathbf{a}_n}_{\text{Previous value}} - \overbrace{s}^{\text{Step size}} \Big[\underbrace{\nabla^2}_{\text{Hessian}} \overbrace{f}^{\text{Target function}}(\mathbf{a}_n)\Big]^{-1} \underbrace{\nabla}_{\text{Gradient}} f(\mathbf{a}_n) \tag{28}$$

Variants

- Scale step size
- Stochastic Newton

Problems

- Large Hessian and inversion expensive
- Slow with a fixed step

$$\mathbf{a}_n - s\,\nabla f(\mathbf{a}_n) \tag{29}$$

Previous value — Step size

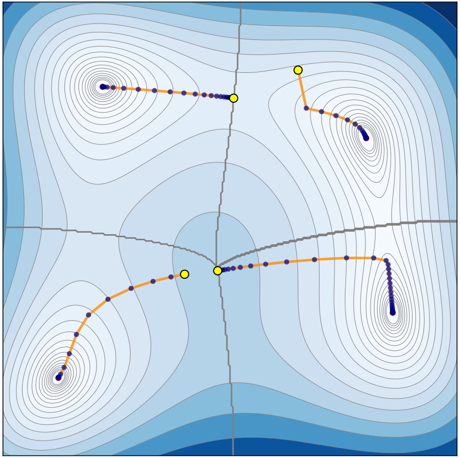
Variants

- Adjust step size
- Line search

Problems

- Slow with fixed step
- Oscillations

Steepest descent



Essentially Newton, but with guessed and updated Hessian

$$\mathbf{p}_{n+1} = -\mathbf{B}_n^{-1} \nabla f(\mathbf{a}_n) \tag{30}$$

Update direction
Approximate Hessian
Previous point

Line search

$$\alpha_{n+1} = \arg\min_{\alpha} f(\mathbf{a}_n + \alpha \mathbf{p}_{n+1}), \qquad \mathbf{s}_{n+1} = \alpha_{n+1}\mathbf{p}_{n+1} \tag{31}$$

Update step length
Actual step

Update

$$\mathbf{a}_{n+1} = \mathbf{a}_n + \mathbf{s}_{n+1}$$

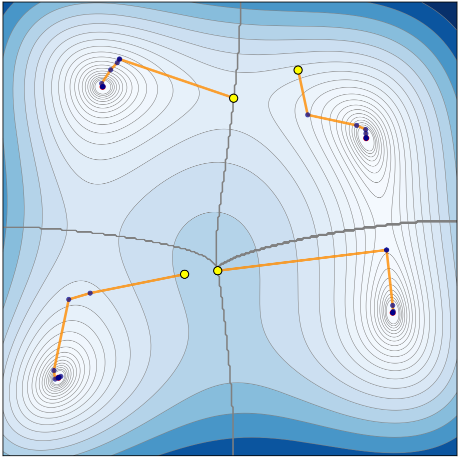
BFGS (continue)

Get gradient response as Hessian approximation

$$\mathsf{y}_{n+1} = \nabla f(\mathsf{a}_{n+1}) - \nabla f(\mathsf{a}_n)$$

Update approximate Hessian

$$\mathsf{B}_{n+1} = \mathsf{B}_n + \frac{\mathsf{y}_{n+1}\mathsf{y}_{n+1}^T}{\mathsf{y}_{n+1}^T s_{n+1}} - \frac{\mathsf{B}_n \mathsf{s}_{n+1} \mathsf{s}_{n+1}^T \mathsf{B}_n^T}{\mathsf{s}_{n+1}^T \mathsf{B}_n \mathsf{s}_{n+1}}$$



Conjugate Gradients (CG)

Avoid oscillations of steepest descent

$$\underbrace{\mathbf{p}_{n+1}}_{\text{Search direction}} = -\nabla f(\mathbf{a}_n) + \underbrace{\beta_n}_{\text{Mixing parameter}} \mathbf{p}_n \qquad \beta_n = \frac{\nabla f(\mathbf{a}_{n+1}^T \nabla f(\mathbf{a}_{n+1}}{\nabla f(\mathbf{a}_n^T \nabla f(\mathbf{a}_n} \tag{32}$$

Search direction

Mixing parameter

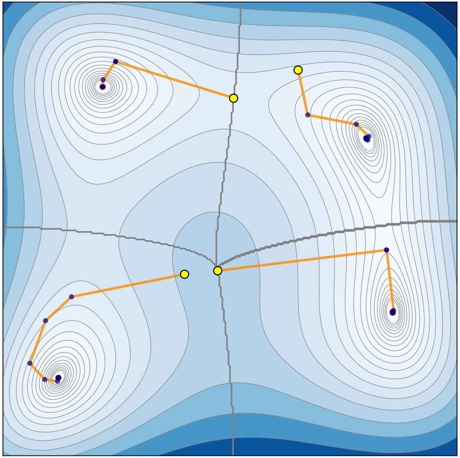
New point found by line search (like BFGS).

Key properties

- Orthogonal search directions
- Quadratic convergence for quadratic functions
- Memory efficient

Problems

- Sensitive to round-off errors
- Requires periodic restarts



Use subset of data for gradient estimation

$$\underbrace{\mathbf{a}_{n+1}}_{\text{Updated parameters}} = \mathbf{a}_n - \underbrace{\eta}_{\text{Learning rate}} \underbrace{\nabla f_i(\mathbf{a}_n)}_{\text{Single sample gradient}} \tag{33}$$

Variants

- Mini-batch SGD
- Learning rate scheduling
- Momentum

Advantages

- Fast iterations
- Escapes local minima

Adaptive moments estimation

Parameter, often 0.9

$$m_n = \beta_1 m_{t-1} + (1 - \beta_1)\nabla f_n \quad (34)$$

Momentum: averaged gradient

Parameter, often 0.999

$$v_n = \beta_2 v_{n-1} + (1 - \beta_2)(\nabla f_n)^2 \quad (35)$$

Scale adjustment

ADAM (continued)

Bias correction, since at start: $m_0 = v_0 = 0$

$$\hat{m}_n = \frac{m_n}{1-\beta_1^n} \quad \hat{v}_n = \frac{v_n}{1-\beta_2^n} \tag{36}$$

Learning rate, often 0.001

$$\mathbf{a}_{n+1} = \mathbf{a}_n - \frac{\alpha \hat{m}_n}{\sqrt{\hat{v}_n} + \epsilon} \tag{37}$$

Small, for stability, often 10^{-8}

Advantages

- Adaptive learning rates
- Robust to hyperparameters

Genetic algorithms

Idea

- Evolutionary optimization approach

Operations

- Selection: Choose fittest individuals
- Crossover: Combine parent solutions
- Mutation: Random modifications
- Replacement: Update population

Parameters

- Population size
- Crossover probability
- Mutation rate
- Selection pressure

Advantages

- Global optimization
- No gradient required
- Handles discontinuous functions

Basin hopping

Idea

- Combine local and global search

Algorithm

- Local minimization
- Random perturbation
- Accept/reject based on energy
- Repeat

Acceptance criterion

$$P = \min\left(1, e^{-\frac{\Delta E}{k_B T}}\right)$$

Advantages

- Escapes local minima
- Uses efficient local methods
- Temperature controls exploration

Applications

- Protein folding
- Molecular conformations
- Glass structure

Optimization: Caveats

Convergence

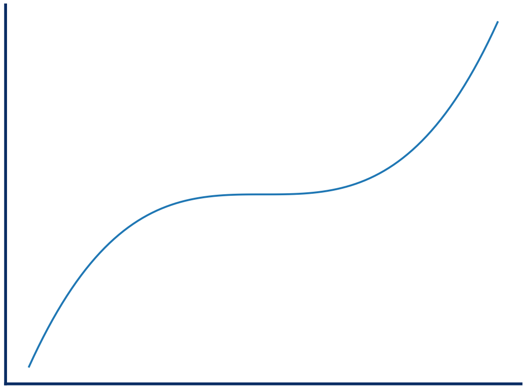
- Hard to establish
- Gradient necessary, but not sufficient
- Hessian expensive
- Local property

Numerical stability

- Finite differences
- Conjugate Gradients
- Shallow minima

Cost of Hessians

- Scales as N^2
- Sometimes only from finite differences



Gradients can be arbitrarily small

Summary

Curse of dimensionality

- Search space quickly increases
- Often forces tiny optimization steps

Preconditioning

- Math not equal to finite-precision implementations
- Transform problem into an equivalent one

Dimensions	Gradients	Hessian	Noise	Minima count	Choice
Few	Yes	Yes	No	Few	Newton
Few/Medium	Yes	No	No	Few	BFGS
Few/Medium	Yes	No	No	Many	Basin hopping
Any	No	No	Any	Many	Genetic algorithm
Large	Yes	No	No	Few	Conjugate gradients
Large	Yes	No	Yes	Few	SGD
Large	Yes	No	Yes	Many	ADAM