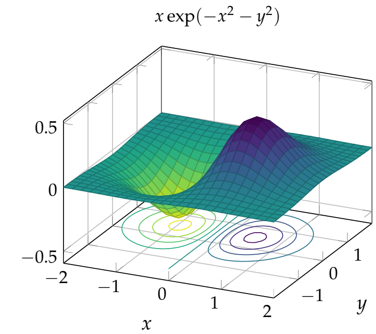
Review: Unconstrained Optimization

Robert Moss — MOSSR@CS.STANFORD.EDU
Stanford University, Stanford, CA 94305 — *AA222/CS361*

$\frac{d}{dx} f(g(x)) = \frac{d}{dx}(f \circ g)(x) = \frac{df}{dg}\frac{dg}{dx}$

$$\nabla^2 f(\mathbf{x}) = \mathbf{H}(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} \end{bmatrix}$$



(2) *Derivatives and Gradients.*

$$f'(x) \approx \overbrace{\frac{f(x+h) - f(x)}{h}}^{\text{forward difference}} \approx \overbrace{\frac{f(x+h/2) - f(x-h/2)}{h}}^{\text{central difference}} \approx \overbrace{\frac{f(x) - f(x-h)}{h}}^{\text{backward difference}}$$

$$f(x) = \text{Re}(f(x+ih)) + h^2\frac{f''(x)}{2!} - \cdots, \qquad f'(x) = \frac{\text{Im}(f(x+ih))}{h} + O(h^2) \text{ as } h \to 0$$

- *Forward accumulation* will auto-diff f with a single forward pass through computational graph (see *dual numbers*: $a + b\epsilon$ where $\epsilon^2 = 0$). *Reverse accumulation* requires n passes.

(3) *Bracketing.*

- *Fibonacci search*, *golden section search* (Fibonacci approximation), *quadratic fit search*, *Shubert-Piyavskii method*,[1] and the *bisection method* (a *root-finding method* expanded by *Brent-Dekker*).

(4) *Local Descent.*

- *Descent direction methods*, (*approximate*) *line search* (learn step size: $\underset{\alpha}{\text{minimize}}\, f(\mathbf{x} + \alpha\mathbf{d})$), *backtracking line search*,[2] *Wolfe conditions*, and *trust regions* (or *restricted step method*).

(5) *First-Order Methods.*

- *Gradient descent*, *conjugate gradient*,[3] *momentum*, *Nesterov momentum*,[4] *adaptive subgradient method* (or *Adagrad*),[5] *RMSProp*,[6] *Adadelta*,[7] *adaptive moment estimation* (or *Adam*),[8] and *hypergradient descent* (augments a `DescentMethod`).

(6) *Second-Order Methods.*

- *Newton's method*[9] (quadratic approximation), *secant method*,[10] and *quasi-Newton methods*.[11]

(7) *Direct Methods.*

- *Cyclic coordinate search*, *Powell's method*,[12] *Hooke-Jeeves*, *generalized pattern search*, *Nelder-Mead simplex method*, and *divided rectangles* (*univariate* and *multivariate DIRECT*).

(8) *Stochastic Methods.*

- *Stochastic gradient descent*, *mesh adaptive direct search*, *simulated annealing* (*Metropolis criterion*), *adaptive simulated annealing*, *cross-entropy method*,[13] *natural evolutionary strategies*, and *covariance matrix adaptation evolutionary strategy* (*CMA-ES*: uses a multivariate Gaussian distribution).

(9) *Population Methods.*

- *Initial population* (uniform, normal, Cauchy), *genetic algorithms* (undergo *crossover* and *mutation*), *differential evolution*, *particle swarm optimization*, *firefly algorithm*, *cuckoo search*, *hybrid methods* (local search using descent methods: *Lamarchian learning* and *Baldwinian learning*).

---

[1] A *global optimization method* where f is *Lipschitz continuous*.

[2] Start big then backtrack.

[3] Overcomes narrow valley issues of *gradient descent*. Its directions are *mutually conjugate* with respect to $\mathbf{A}$. Approximations of β are *Fletcher-Reeves* and *Polak-Ribière*.

[4] Reduce overshooting at bottom.

[5] Dulls high gradients, increases influence of infreq. updated params.

[6] Extends *Adagrad* to avoid effects of a monotonically decreasing learning rate.

[7] Overcomes *Adagrad*'s monotonically decreasing learning rate by eliminating α in favor of:

$$x' = x - \frac{\text{RMS}(\Delta x)}{\epsilon + \text{RMS}(g)} g$$

[8] Adapts learning rate to each parameter, biases then corrects decay in momentum $\mathbf{v}$ and sq. gradient $\mathbf{s}$.

[9]
$$x' = x - \frac{f'(x)}{f''(x)}$$

[10] Unlike *Newton's method*, estimates f'', only requires f':

$$f''(x) \approx \frac{f'(x) - f'(x^{(k-1)})}{x - x^{(k-1)}}$$

[11] Inverse Hessian $\mathbf{H}^{-1}$ approximation: *DFP, BFGS, L-BFGS* approximate line search.

[12] Non-orthogonal directions.

[13] Sample, select, fit, repeat.