Exercise Sheet 5

Optimization and Integration

1. Newton's Method in $\mathbb{R}^2$

Consider the function

$$f(x, y) = x^2 + xy + y^2 - 4x - 2y.$$

(i) Compute the gradient $\nabla f(x, y)$ and the Hessian $H(x, y) = \nabla^2 f(x, y)$.

(ii) Starting from $(x_0, y_0) = (0, 0)$, perform one Newton step and state the resulting position.

(iii) How does Newton's method behave for this function? Are there any instabilities? How many steps will it require to converge?

Solution:

(i) The gradient is:

$$\nabla f(x, y) = \begin{pmatrix} 2x + y - 4 \\ x + 2y - 2 \end{pmatrix}.$$

The Hessian is:

$$H(x, y) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

(ii) At $(0, 0)$, the Newton system $H\mathbf{p} = -\nabla f$ is:

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \end{pmatrix}.$$

Solving gives $\mathbf{p} = (2, 0)^T$, so $(x_1, y_1) = (2, 0)$.

(iii) Since f is quadratic, Newton's method converges in exactly one step from any starting point (the quadratic Taylor approximation is exact).

2. Gradient Descent Stability

Consider the function

$$f(x, y) = \frac{1}{2}(x^2 + 4y^2).$$

(i) Write out the gradient descent iteration with step size η.

(ii) Express the updates for x and y separately.

(iii) For what range of η does gradient descent converge to the minimum?

(iv) Besides faster convergence, state an advantage of second-order methods (e.g., Newton's method) over first-order methods (e.g., gradient descent).

(v) Gradient descent with momentum introduces a velocity $\mathbf{v}_k$ that accumulates gradient information:

$$\mathbf{v}_{k+1} = \beta \mathbf{v}_k + \eta \nabla f(\mathbf{x}_k), \qquad \mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{v}_{k+1}.$$

If the gradient were constant, what effective step size does momentum achieve in the long run? How much larger is it than η when $\beta = 0.999$?

Solution:

(i) The gradient is $\nabla f = (x, 4y)^T$. Gradient descent:

$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} = \begin{pmatrix} x_k \\ y_k \end{pmatrix} - \eta \begin{pmatrix} x_k \\ 4y_k \end{pmatrix}.$$

(ii) Separating components: $x_{k+1} = (1-\eta)x_k$ and $y_{k+1} = (1-4\eta)y_k$.

(iii) For convergence, both iterations must be stable: $|1-\eta| < 1$ and $|1-4\eta| < 1$. The intersection of these constraints gives $0 < \eta < \frac{1}{2}$.

(iv) Second-order methods automatically determine an appropriate step size using curvature information, whereas first-order methods require manual tuning of η.

(v) For a constant gradient $\mathbf{g}$, the velocity converges to $\mathbf{v}_\infty = \beta \mathbf{v}_\infty + \eta \mathbf{g}$, giving $\mathbf{v}_\infty = \frac{\eta}{1-\beta}\mathbf{g}$. The effective step size is $\frac{\eta}{1-\beta}$. For $\beta = 0.999$, this is $\frac{\eta}{0.001} = 1000\eta$—a $1000\times$ acceleration.

3. Numerical Integration

Let $f(x) = 6x(1-x)$ and consider the integral

$$I = \int_0^1 f(x)\, dx.$$

(i) Compute I analytically.

(ii) Approximate I using the trapezoid rule with $N = 3$ quadrature points, and compute the absolute error.

(iii) Approximate I using Monte Carlo integration with $N = 3$ samples $X_i \sim \text{Uniform}[0,1]$. What is the standard deviation of this estimator? You may use $\int_0^1 f(x)^2\, dx = \frac{6}{5}$.

(iv) State two characteristic differences between quadrature and Monte Carlo integration when integrating over the d-dimensional unit cube $[0,1]^d$.

Solution:

(i) $I = \int_0^1 (6x - 6x^2)\, dx = \left[3x^2 - 2x^3\right]_0^1 = 3 - 2 = 1.$

(ii) With $N = 3$ points: $x_0 = 0$, $x_1 = \frac{1}{2}$, $x_2 = 1$, and $h = \frac{1}{2}$.

The function values are $f(0) = 0$, $f(\frac{1}{2}) = \frac{3}{2}$, $f(1) = 0$.

Trapezoid rule:

$$I_{\text{trap}} = \frac{h}{2}\big[f(x_0) + 2f(x_1) + f(x_2)\big] = \frac{1}{4}\,[0 + 3 + 0] = \frac{3}{4}.$$

Absolute error: $\left|\frac{3}{4} - 1\right| = \frac{1}{4}$.

(iii) The Monte Carlo estimator is $\hat{I} = \frac{1}{N}\sum_{i=1}^N f(X_i)$.

Its standard deviation is $\sigma_{\hat{I}} = \frac{\sigma_f}{\sqrt{N}}$, where $\sigma_f^2 = \int_0^1 f(x)^2\, dx - I^2 = \frac{6}{5} - 1 = \frac{1}{5}$.

Thus $\sigma_{\hat{I}} = \frac{1}{\sqrt{5 \cdot 3}} = \frac{1}{\sqrt{15}}$.

(iv) Two characteristic differences:

a) **Determinism:** Quadrature is deterministic and gives the same error every time. Monte Carlo is stochastic and in principle gives a different answer on each run (unless the random number generator is re-seeded).

b) **Curse of dimensionality:** In d dimensions, quadrature with N points per axis requires N^d evaluations and suffers from the curse of dimensionality. Monte Carlo has error $O(1/\sqrt{N})$ regardless of dimension, making it preferable for high-dimensional integrals.