UNIVERSITY COMPUTER SCIENCE • EXERCISE SHEET

13.5 Optimisation and learning

Name: _____ Class: _____ Date: _____ **Total: 24 marks**

KEY WORDS gradient descent • learning rate

Objective

Run gradient descent by hand and respect the stability bound.

1 Key knowledge and worked examples

■ **The rule**

$w_{t+1} = w_t - \eta \nabla L(w_t)$ ✓. For $L(w) = (w-3)^2$: $w_{t+1} = w_t - 2\eta(w_t - 3)$, stable iff $\eta < 1$ ✓.

■ **Worked example: three rates**

$\eta = 0.1$: 0.6, 1.08, 1.464, 1.771, 2.017 (converging). $\eta = 1$: 0, 6, 0, 6 (oscillating). $\eta = 0.6$: 3.6, 2.88, 3.024 (damped) ✓.

2 Practice

2.1 For $L(w) = (w-3)^2$ with $\eta = 0.1$ from $w_0 = 4$: give w_1, w_2. [3]

2.2 What is the stability bound on η here? [3]

2.3 What shape does $L = (x_1 - 1)^2 + 100(x_2 - 1)^2$ have? [3]

3 Exam-style questions

3.1 Run gradient descent on $L(w) = (w+2)^2$ (note the sign) from $w_0 = 1$ with $\eta = 0.25$ for four steps.

Give the update rule specialised to this loss. List every iterate (exact fractions or 4 decimals). State the limit and whether the steps overshoot. Give the iterate sequence for $\eta = 0.75$ and state its qualitative difference. [8]

3.2 Explain, with numbers, why feature scaling changes the usable learning rate.

For $L = (x_1 - 1)^2 + 100(x_2 - 1)^2$ from $(0, 0)$: state the curvature along each axis and the stability bound on η for a single shared rate. Give the number of steps needed to reduce each coordinate's error by half under that bound. State what standardising the second feature does to the bound and to the descent. [7]

Solutions

2.1

- $w_1 = 4 - 0.2(1) = 3.8$; $w_2 = 3.8 - 0.2(0.8) = 3.64$ ✓.

2.2

- $|1 - 2\eta| < 1 \iff 0 < \eta < 1$ ✓.

2.3

- A ravine: 100x curvature mismatch between the axes ✓.

3.1

- Update: $w_{t+1} = w_t - 0.5(w_t + 2) = 0.5w_t - 1$ ✓.
- Iterates: $1 \to -0.5 \to -1.25 \to -1.625 \to -1.8125$ ✓.
- Limit -2 (the minimum); errors shrink monotonically, no overshoot ✓.
- With $\eta = 0.75$: $w_{t+1} = w_t - 1.5(w_t + 2) = -0.5w_t - 3$: $1 \to -3.5 \to -1.25 \to -2.375 \to -1.8125$ – alternating sign of the error, damped oscillation ✓.

3.2

- Curvatures: 2 and 200 (the loss's second derivatives) ✓.
- Shared-rate stability: $\eta < 2/200 = 0.01$ (the tighter axis rules) ✓.
- Halving the error per coordinate: the slow axis contracts at $|1 - 2\eta| = 0.98$ per step – about 34 steps to halve ✓.
- Standardising the second feature equalises curvature at 2: the bound becomes $\eta < 1$ and one rate serves both axes ✓.

exercises.lol exercises.com.cn