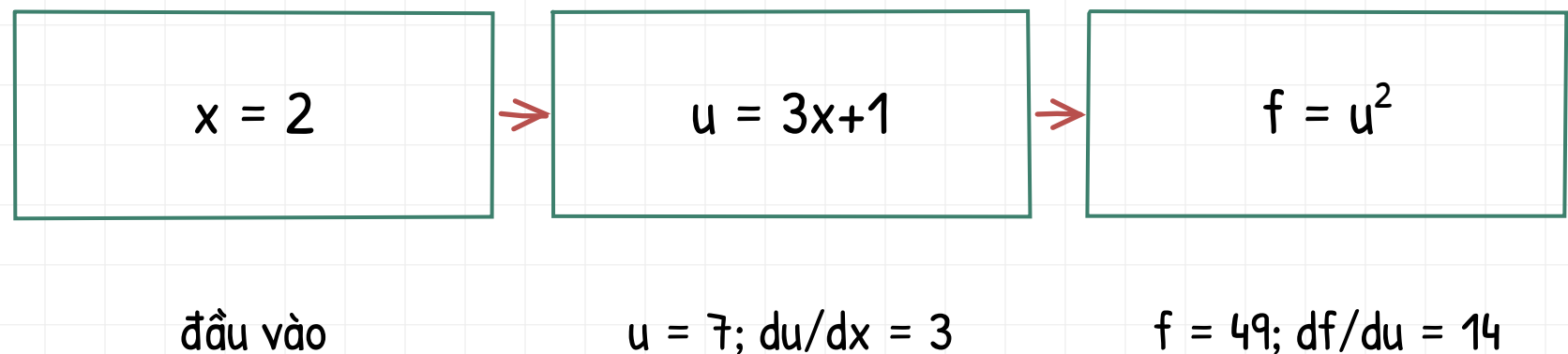


Nhân độ nhạy qua từng chặng

x nhích một chút; chặng đầu khuếch đại 3 lần, chặng sau 14 lần.



Từ hình đến công thức

$$\Delta u \approx (du/dx) \cdot \Delta x; \Delta f \approx (df/du) \cdot \Delta u.$$

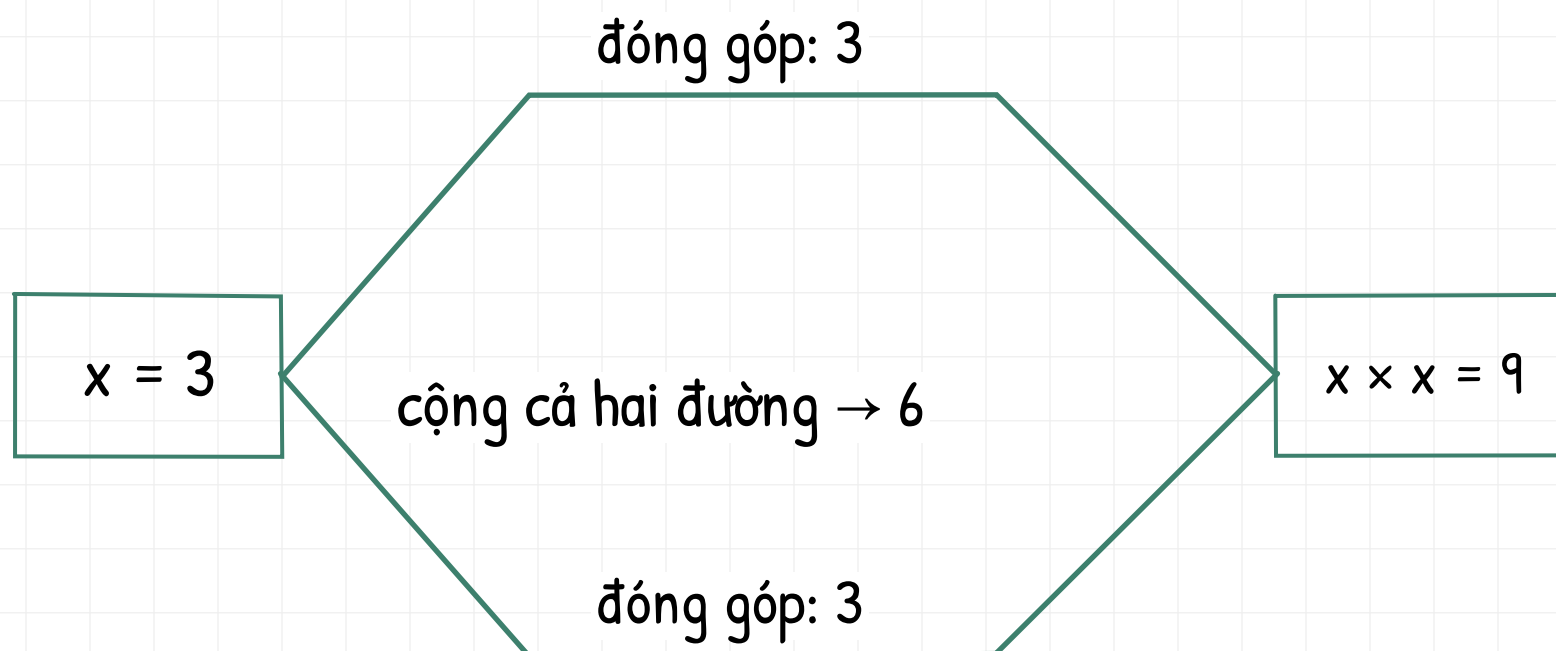
Ghép hai bước: $df/dx = (df/du) \cdot (du/dx)$.

$f = (3x+1)^2$ tại $x=2$: $df/dx = 14 \cdot 3 = 42$. Quên nhân 3 là bỏ mất chặng đầu.

Chain rule là ghép các phép tuyến tính hoá tại chỗ.

Một biến, hai đường đi

x xuất hiện hai lần trong $x \cdot x$: ảnh hưởng phải được cộng lại.



Từ hình đến công thức

Tại $x=3$: nhánh trái góp 3, nhánh phải góp 3 $\rightarrow$ tổng gradient là 6.

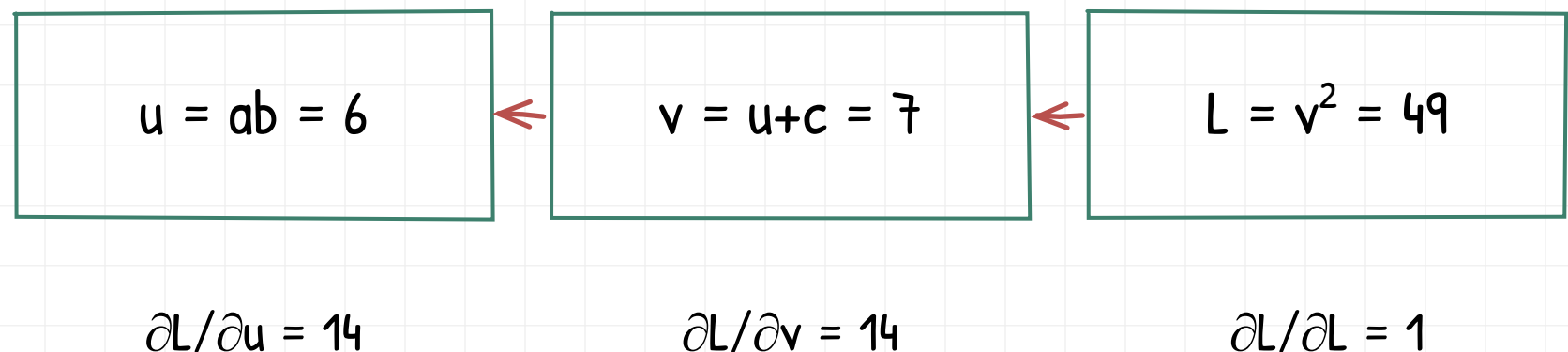
Tổng quát: $\partial L / \partial x = \sum_v (\partial L / \partial v) \cdot (\partial v / \partial x)$, với v là các nút dùng x .

Trong backward dùng $\text{grad} += \text{đóng_góp}$; $\text{gán} =$ sẽ ghi đè nhánh trước.

Nhân dọc mỗi đường; cộng giữa những đường hội lại.

Backprop chạy ngược đồ thị

Lưu các giá trị khi chạy xuôi; dùng chúng để lan gradient ngược.



Từ hình đến công thức

Với $a=3$, $b=2$, $c=1$: $\partial L / \partial a = 28$, $\partial L / \partial b = 42$, $\partial L / \partial c = 14$.

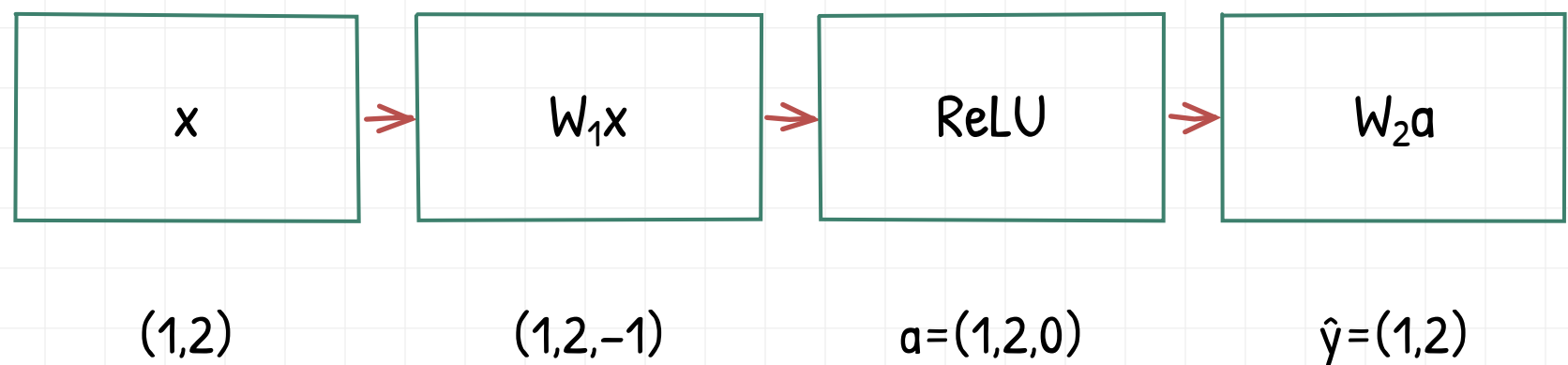
Forward $\rightarrow$ đặt grad đầu ra $= 1 \rightarrow$ duyệt topo ngược $\rightarrow$ cộng các đóng góp.

Đặt gradient về 0 trước lượt backward mới, trừ khi chủ ý tích lũy.

Backward phải chờ đủ đóng góp từ các nút phía sau.

Một mạng nhỏ, chạy xuôi

$x=(1,2)$, nhãn $y=(0,1)$. Lớp ẩn có một đầu vào âm bị ReLU chặn.



Từ hình đến công thức

W_1 có các hàng $(1,0)$, $(0,1)$, $(1,-1)$. W_2 có các hàng $(1,0,1)$, $(0,1,0)$.

$r=\hat{y}-y=(1,1)$; $L=\text{mean}(r^2)=1$.

Lấy trung bình trên $m=2$ phần tử: $\partial L/\partial \hat{y}=2r/m=(1,1)$.

Mean và sum khác hệ số gradient; phải khớp cách định nghĩa loss.

Chạy ngược qua ma trận

Gradient có cùng kích thước với ma trận trọng số tương ứng.

$$\begin{array}{c} \partial L / \partial W_2 \\ \left[\begin{array}{cc} 1 & 2 \\ 1 & 2 \end{array} \right] \end{array} \longrightarrow \begin{array}{c} \partial L / \partial W_1 \\ \left[\begin{array}{cc} 1 & 2 \\ 1 & 2 \\ 0 & 0 \end{array} \right] \end{array}$$

hàng cuối = 0: cổng ReLU đang đóng

Từ hình đến công thức

$$g = \partial L / \partial \hat{y} = (1, 1); \quad \partial L / \partial W_2 = g \cdot a^T.$$

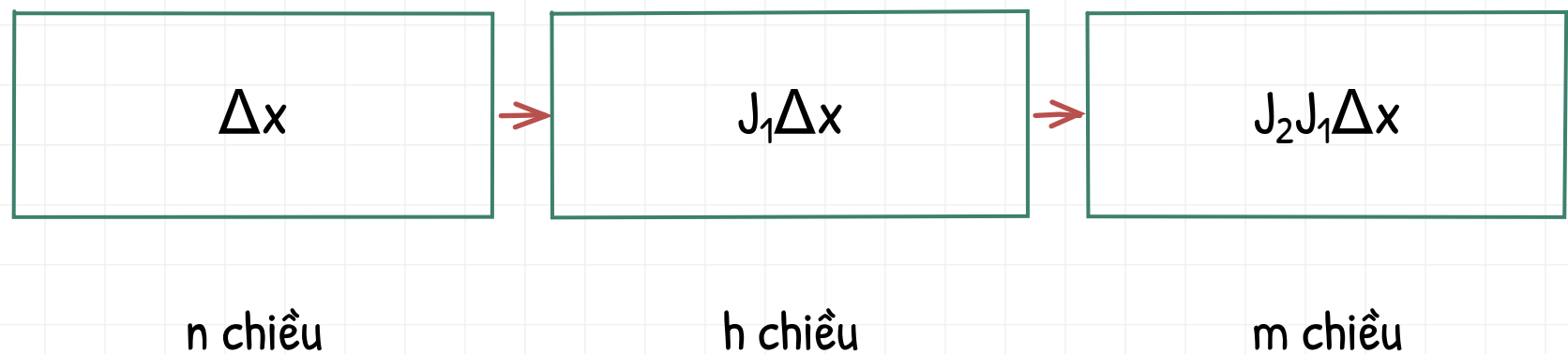
$$\text{Về lớp ẩn: } g_z = (W_2^T g) \odot 1[z > 0] = (1, 1, 0).$$

$\partial L / \partial W_1 = g_z \cdot x^T$. Hàng 3 bằng 0 cho mẫu này; mẫu khác có thể mở lại cổng.

Tích ngoài cập nhật trọng số; chuyển vị đưa gradient về đầu vào.

Jacobian ghép như ma trận

Một nhiễu nhỏ đi xuôi qua các lớp; gradient loss đi theo chiều ngược.



Từ hình đến công thức

$J[i,j] = \partial F_i / \partial x_j$; J có m hàng đầu ra, n cột đầu vào.

$J_{\text{tổng}} = J_2 J_1$; gradient dạng cột: $\nabla_x L = J_1^T J_2^T \nabla_y L$.

Mạng ở trang trước: $J = W_2 \cdot \text{diag}(1,1,0) \cdot W_1 = I_2$ tại $x = (1,2)$.

Thứ tự nhân có ý nghĩa; chỉ đổi thứ tự đặt ngoặc khi hợp lệ.

Tái dùng thay vì khai triển

Cùng một đồ thị được đi hai lượt; không cần viết đạo hàm khổng lồ.

Cách	Một đầu ra, n tham số
Sai phân trung tâm	$2n$ lần tính hàm
Reverse mode	1 forward + 1 backward
Bộ nhớ	lưu / tính lại giá trị trung gian

Từ hình đến công thức

Reverse mode tính vector-Jacobian product, không dựng toàn bộ Jacobian của mạng.

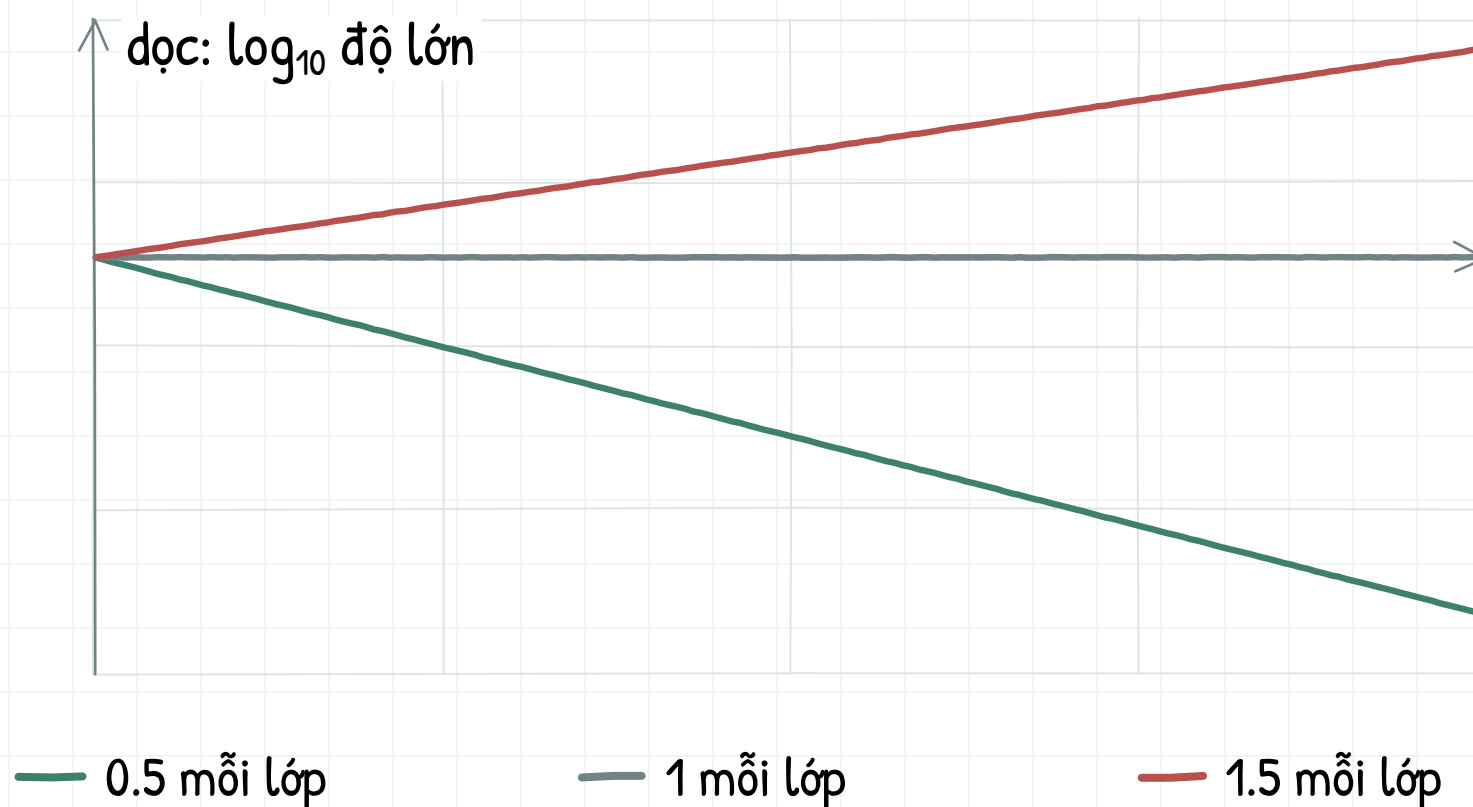
Chi phí backward thường cùng bậc với forward; không phải đúng hai forward.

Forward mode tiện khi ít đầu vào; reverse mode tiện khi ít đầu ra.

Hiệu quả đến từ thứ tự tính và dùng lại kết quả trung gian.

Nhân nhiều lần: tắt hoặc nổ

Mô hình vô hướng: mỗi lớp nhân gradient với cùng một hệ số r .



Từ hình đến công thức

Sau 20 lớp: $0.5^{20} \approx 9.537 \times 10^{-7}$; $1.5^{20} \approx 3325.26$.

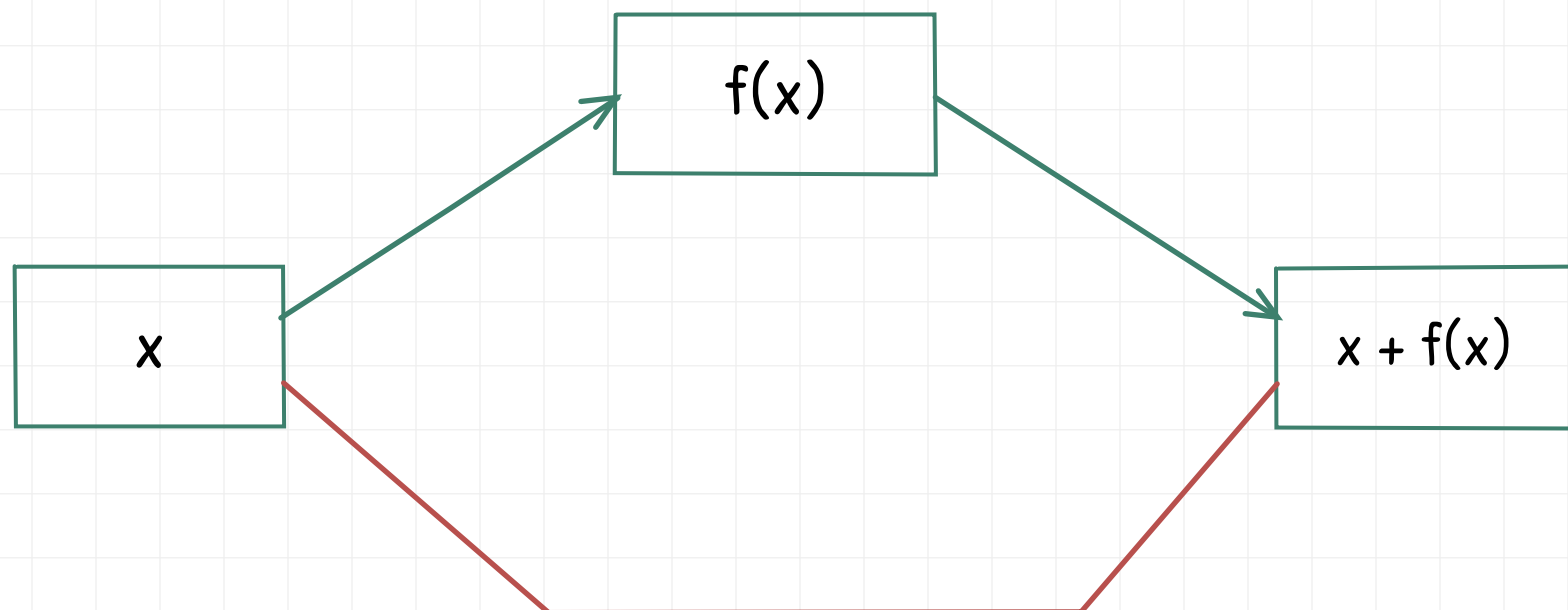
Mạng thật nhân nhiều Jacobian khác nhau. Hướng và các giá trị suy biến đều quan trọng.

Không thể chỉ nhìn một eigenvalue của từng lớp để kết luận tích sẽ ổn định.

Đây là phép nhân lặp của chain rule, không phải một lỗi riêng của loss.

Residual thêm đường đi thẳng

Nhánh x bỏ qua $f(x)$, rồi hai nhánh cộng lại ở đầu ra.



nhánh đồng nhất: đạo hàm = 1

Từ hình đến công thức

$h(x) = x + f(x) \rightarrow J_h = I + J_f$. Nhánh f nhỏ thì khối gần đồng nhất.

Vô hướng, $f' = 0.05$: $(1 + f')^{20} \approx 2.6533$; $f' = 0.5$ vẫn cho 3325.26.

Có đường identity không bảo đảm gradient khác 0: $f' = -1$ làm $1 + f' = 0$.

Residual hỗ trợ dòng gradient; vẫn cần khởi tạo và kiểm độ lớn.

Tự kiểm trước khi đi tiếp

Che đáp án, thử nói bằng lời

1. $(3x+1)^2$ tại $x=2$ có đạo hàm bao nhiêu?
2. $x \cdot x$ tại $x=3$: vì sao phải $+=$?
3. Loss là mean trên m phần tử: gradient MSE?
4. Kiểm backward tự viết thế nào?
5. Residual có bảo đảm không mất gradient?

Đáp nhanh

1. $42 = 14 \cdot 3$.
2. Hai đường mỗi đường góp 3; tổng đúng là 6.
3. $2(\hat{y}-y)/m$; thiếu $/m$ làm gradient lớn gấp m .
4. So sai phân trung tâm nhiều h, tránh góc ReLU; xét cả sai số tương đối.
5. Không. $l+J_f$ vẫn có thể co, nổ hoặc triệt tiêu.