

تکلیف اول درس شبکه های عصبی

وحید ملکی
شماره دانشجویی: ۴۰۳۱۳۰۰۴

۱۴۰۴ آبان ۲۶

سوال ۳

الف: رسم توابع فعال‌سازی و مشتق آن‌ها

با استفاده از کتابخانه matplotlib، تابع فعال‌سازی زیر به همراه مشتق آن‌ها در بازه $[5, -5]$ رسم شده‌اند:

ReLU •

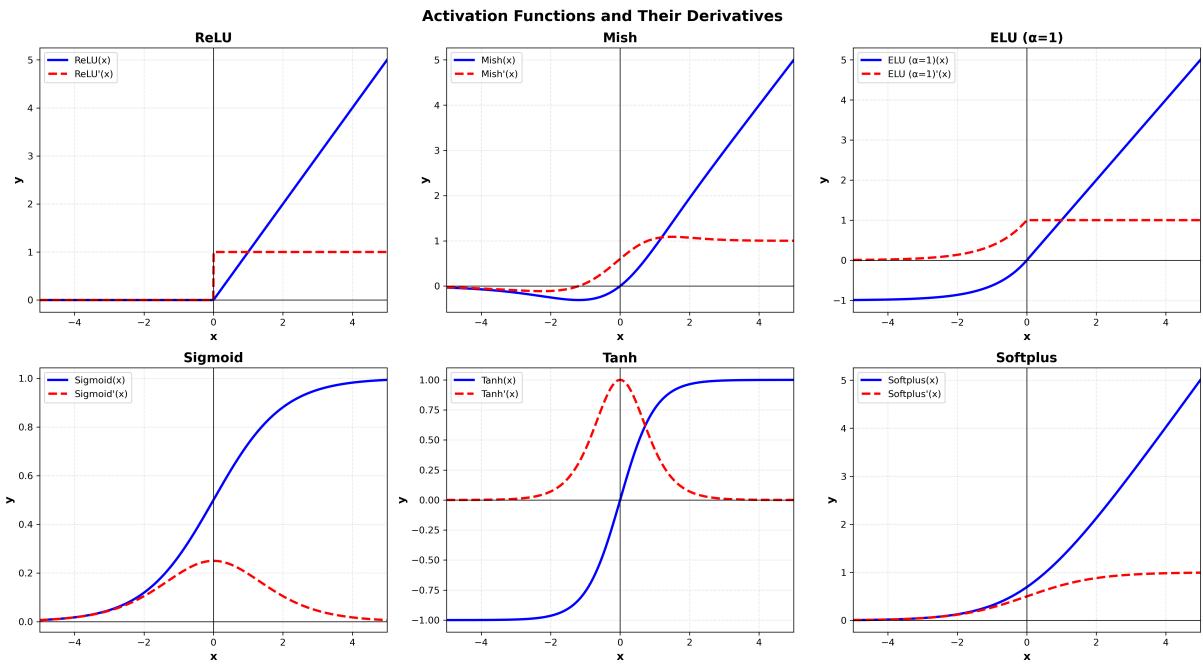
Mish •

$(\alpha = 1)$ ELU •

Sigmoid •

Tanh •

Softplus •



شکل ۱: توابع فعال‌سازی (خط آبی پرنگ) و مشتق آنها (خط قرمز نقطه‌چین) برای ReLU، Mish، ELU، Sigmoid، Tanh، Softplus و $\alpha=1$ ELU.

مشاهدات از شکل ۱

ReLU: برای $x > 0$ خطی با شیب ۱، برای $x \leq 0$ صفر. مشتق: ۱ در ناحیه مثبت، ۰ در ناحیه منفی.

Mish: در ناحیه منفی $f(x) = x \cdot \tanh(\ln(1+e^x))$ نرم تر از ReLU. مشتق همیشه مثبت و نرم است.

ELU: برای $x > 0$ همانند ReLU، برای $x \leq 0$ به صورت ثابی به α نزدیک می‌شود. مشتق در ناحیه منفی از مرگ نورون جلوگیری می‌کند.

Sigmoid: خروجی در $[0, 1]$ ، مشتق حداکثر ۲۵۰۰ در $x = 0$ و در $|x|$ بزرگ به صفر نزدیک می‌شود (اشباع).

Tanh: برای $x = 0$ خروجی در $[-1, 1]$ ، متقارن حول صفر، مشتق حداکثر ۱ در $x = 0$.

Softplus: نسخه نرم ReLU. مشتق آن دقیقاً تابع سیگموید است.

ب: مشکل اشباع شدن (Vanishing Gradient) در تابع Sigmoid

وقتی $|x|$ بزرگ باشد، خروجی سیگموید به ۱ یا ۰ نزدیک شده و مشتق آن تقریباً صفر می‌شود:

$$\sigma'(x) = \sigma(x)(1 - \sigma(x)) \leq 0.25$$

در شبکه‌های عمیق، گرادیان در هر لایه در مشتق تابع فعال‌سازی ضرب می‌شود. اگر چندین لایه مشتق نزدیک به صفر داشته باشد، گرادیان در لایه‌های اولیه ناپدید (vanish) می‌شود و وزن‌ها به روزرسانی نمی‌شوند. آیا جایگزینی با Tanh این مشکل را کاملاً بطرف می‌کند؟

خیر، اما به طور قابل توجهی کاهش می‌یابد.
مشتق \tanh' :

$$\tanh'(x) = 1 - \tanh^2(x) \leq 1$$

حداکثر مشتق آن ۱ است (۴ برابر سیگموید). در نواحی اشباع همچنان به صفر میل می‌کند، پس مشکل کاملاً حذف نمی‌شود، اما در عمل گرادیان‌ها کمتر ناپدید می‌شوند.

ج: چرا Tanh بهینه‌سازی را نسبت به Sigmoid آسان‌تر می‌کند؟

۱. مشتق بزرگ‌تر: خداکثر مشتق \tanh برابر ۱ است (در مقابل ۲۵۰۰ برابر سیگموید) ۰ گرادیان‌های بزرگ‌تر همگرایی سریع‌تر.

۲. خروجی مرکز حول صفر: (Zero-Centered) خروجی \tanh در $[-1, 1]$ و میانگین نزدیک صفر است، در حالی که سیگموید در $[0, 1]$ و میانگین ۰.۵ است. خروجی همیشه مثبت در سیگموید باعث بهروزرسانی زیگزاگی وزن‌ها می‌شود، اما \tanh اجازه جهت‌های مختلف گرادیان را می‌دهد ۰ بهینه‌سازی روان‌تر.

۳. تقارن حول مبدأ: \tanh تابع فرد است $\tanh(-x) = -\tanh(x)$ ۰ یادگیری الگوهای متقارن آسان‌تر.

جدول ۱: مقایسه تابع فعال‌سازی Sigmoid و Tanh

Tanh	Sigmoid	ویژگی
$[-1, 1]$	$[0, 1]$	بازه خروجی
بله	خیر	مرکز حول صفر؟
۱	۲۵۰۰	حداکثر مشتق
بسیار شدید	متوسط	شدت Gradient Vanishing
سریع‌تر	کند	سرعت همگرایی

در نتیجه، \tanh در دهه ۲۰۱۰ جایگزین اصلی sigmoid در لایه‌های مخفی بود، اما امروزه ReLU و خانواده آن به دلیل عدم Galleria اشباع در ناحیه مثبت، استاندارد هستند.