CPEN 455: Deep Learning — HW1 Tutorial

Presented by: Jia Jun Cheng Xian

October 6, 2025

Setup: Single-hidden-layer MLP with Dropout

Model (hidden pre-activation and activation):

$$h = \sigma(Wx + b) \in \mathbb{R}^{M\times 1}, \quad \sigma = \mathrm{ReLU}$$

Dropout during training:

$$\tilde{h} = \frac{m}{1-p} \odot h, \qquad m[i] \sim \mathrm{Bernoulli}(1-p) \text{ i.i.d.}$$

At test time: use $\tilde{h} = h$ (no masking).

- $\odot$ denotes elementwise (Hadamard) product.
- p is the drop probability; $1 - p$ is the keep probability.

1.1 Why scale by $1/(1-p)$?

Goal: Keep the layer's *expected* output magnitude the same between train and test.

$$\mathbb{E}[\tilde{h}] = \mathbb{E}\left[\frac{m}{1-p} \odot h\right] = \frac{\mathbb{E}[m]}{1-p} \odot \mathbb{E}[h] = \frac{1-p}{1-p}\mathbb{E}[h] = \mathbb{E}[h].$$

- Without the factor, $\mathbb{E}[m \odot h] = (1-p)\mathbb{E}[h] \neq \mathbb{E}[h]$.
- Scaling by $1/(1-p)$ makes train-time expectation match test-time expectation.

1.2 Variance of h (before Dropout) under given assumptions

Assumptions: $x \sim \mathcal{N}(0, I)$, $b = 0$, $WW^\top = I_M$, $\sigma = \mathrm{ReLU}$.

- Then $z := Wx + b \sim \mathcal{N}(0, I_M)$ and $h[i] = \max\{0, z[i]\}$.
- For $z \sim \mathcal{N}(0, 1)$:

$$\mathbb{E}[h[i]] = \mathbb{E}[\max(0, z)] = \frac{1}{\sqrt{2\pi}}, \qquad \mathbb{E}[h[i]^2] = \frac{1}{2}.$$

  Hence

$$\mathrm{Var}[h[i]] = \mathbb{E}[h[i]^2] - \mathbb{E}[h[i]]^2 = \frac{1}{2} - \frac{1}{2\pi}.$$

- Since coordinates are independent here, $\mathrm{Var}[h] = \mathrm{diag}\left(\frac{1}{2} - \frac{1}{2\pi}\right)$.

1.2 Variance after Dropout: $\tilde{h} = \frac{m}{1-p} \odot h$

Key facts: $m[i] \in \{0,1\}$, $\mathbb{E}[m[i]] = 1-p$, $\mathbb{E}[m[i]^2] = 1-p$, and m is independent of h.

$$
\begin{aligned}
\mathrm{Var}[\tilde{h}[i]] &= \frac{1}{(1-p)^2}\,\mathrm{Var}[m[i]\,h[i]] \\
&= \frac{1}{(1-p)^2}\Big(\mathbb{E}[m[i]^2]\mathbb{E}[h[i]^2] - \mathbb{E}[m[i]]^2\,\mathbb{E}[h[i]]^2\Big) \\
&= \frac{1}{(1-p)^2}\Big((1-p)\cdot\tfrac{1}{2} - (1-p)^2\cdot\tfrac{1}{2\pi}\Big) \\
&= \frac{1}{2(1-p)} - \frac{1}{2\pi}.
\end{aligned}
$$

Matrix form: $\mathrm{Var}[\tilde{h}] = \mathrm{diag}\left(\frac{1}{2(1-p)} - \frac{1}{2\pi}\right).$

1.3 How many units are kept?

Each unit is kept i.i.d. with probability $1 - p$:

$$K := \sum_{i=1}^{M} \mathbf{1}\{m[i] = 1\} \sim \mathrm{Binomial}(M,\, 1-p)\,.$$

- Expectation: $\mathbb{E}[K] = M(1 - p)$.
- PMF: $\Pr(K = k) = \binom{M}{k}(1-p)^k p^{M-k}$ for $k = 0, 1, \ldots, M$.

1.4 Poisson limit (large M, rare keep)

Regime: $M \to \infty$, $1 - p \to 0$ with $\lambda := M(1-p)$ fixed.

- Then the binomial $K \sim \mathrm{Binomial}(M, 1-p)$ converges in distribution to

$$K \xrightarrow{d} \mathrm{Poisson}(\lambda), \qquad \Pr(K = k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

- Intuition: many trials, very small keep-probability $\Rightarrow$ rare-event process.

1.5 Random width $M \sim \mathrm{Poisson}(\lambda)$ (thinning)

Setup: Draw $M \sim \mathrm{Poisson}(\lambda)$ units, then keep each independently with prob. $1 - p$.

- By *Poisson thinning*, the kept-count is

$$K \sim \mathrm{Poisson}\big(\lambda(1 - p)\big).$$

- Proof sketch: condition on M, $K|M \sim \mathrm{Binomial}(M, 1 - p)$; marginalizing over M yields Poisson with mean $\lambda(1 - p)$.

Key Takeaways (Q1)

- Scaling by $1/(1-p)$ preserves the *expected* activation at train time.
- Under Gaussian+ReLU assumptions, $\mathrm{Var}[h[i]] = \frac{1}{2} - \frac{1}{2\pi}$ and $\mathrm{Var}[\tilde{h}[i]] = \frac{1}{2(1-p)} - \frac{1}{2\pi}$.
- Kept-unit count is $\mathrm{Binomial}(M, 1-p)$; admits Poisson limit and Poisson thinning variants.

Setup: Single hidden layer with batch inputs

Inputs: $X \in \mathbb{R}^{B \times N}$ (rows are samples)
Pre-activations: $Y = XW^\top + b^\top \in \mathbb{R}^{B \times M}$
Activation: $H = \sigma(Y)$ with ReLU $\sigma(u) = \max(0, u)$

Broadcasting: $XW^\top \in \mathbb{R}^{B \times M}$, $b^\top \in \mathbb{R}^{1 \times M}$, so $Y = XW^\top + b^\top$ adds row-wise.

Batch Normalization on Y:

$$\mu[j] = \frac{1}{B} \sum_{i=1}^{B} Y[i,j], \qquad v[j] = \frac{1}{B} \sum_{i=1}^{B} \left(Y[i,j] - \mu[j]\right)^2,$$

$$\hat{Z}[i,j] = \frac{Y[i,j] - \mu[j]}{\sqrt{v[j] + \varepsilon}}, \qquad \hat{Y}[i,j] = \gamma[j]\hat{Z}[i,j] + \beta[j],$$

with learnable $\gamma, \beta \in \mathbb{R}^{M \times 1}$ and small $\varepsilon > 0$.

2.1 Why do we need ε?

- Numerical stability: protects division by 0 or very small $v[j]$ when a feature is (near) constant in a mini-batch.
- Stabilizes gradients (denominator $\sqrt{v[j]+\varepsilon}$ bounded away from 0), preventing exploding updates.
- Has no effect asymptotically when $v[j] \gg \varepsilon$; typically $\varepsilon \in [10^{-5}, 10^{-3}]$.

2.2 Mean and variance of $\hat{Y}$ (ignore ε for this part)

Define $Z[i,j] = \dfrac{Y[i,j] - \mu[j]}{\sqrt{v[j]}}$. By construction:

$$\mathbb{E}[Z[i,j]] = 0, \qquad \mathrm{Var}[Z[i,j]] = 1.$$

Since $\hat{Y}[i,j] = \gamma[j]\, Z[i,j] + \beta[j]$ is an affine transform,

$$\mathbb{E}[\hat{Y}[i,j]] = \beta[j], \qquad \mathrm{Var}(\hat{Y}[i,j]) = \gamma[j]^2.$$

Takeaway: BN recenters to β and rescales variance to γ^2 (per feature).

2.3 Backprop: overview of the computation graph

$$X \longrightarrow Y = XW^\top + b^\top \longrightarrow \hat{Y} = \mathrm{BN}(Y; \gamma, \beta) \longrightarrow H = \mathrm{ReLU}(\hat{Y}) \longrightarrow \ell(H)$$

Given upstream gradient $\frac{\partial \ell}{\partial H} \in \mathbb{R}^{B \times M}$, we backprop:

$$\frac{\partial \ell}{\partial \hat{Y}} = \frac{\partial \ell}{\partial H} \odot \mathbf{1}\{\hat{Y} > 0\}.$$

2.3 Backprop through BN: parameter gradients

Work featurewise ($j = 1, \ldots, M$). Let $g[i,j] = \frac{\partial \ell}{\partial \hat{Y}[i,j]}$.

$$\frac{\partial \ell}{\partial \beta[j]} = \sum_{i=1}^{B} g[i,j],$$

$$\frac{\partial \ell}{\partial \gamma[j]} = \sum_{i=1}^{B} g[i,j]\; \hat{Z}[i,j].$$

Define $g_Z[i,j] = g[i,j]\;\gamma[j]$ and $\mathrm{std}[j] = \sqrt{v[j] + \varepsilon}$ for brevity.

2.3 Backprop through BN: input gradients

Per-feature scalar form (for fixed j):

$$\frac{\partial \ell}{\partial v[j]} = \sum_{i=1}^{B} \frac{\partial \ell}{\partial \hat{Y}[i,j]} \frac{\partial \hat{Y}[i,j]}{\partial v[j]} = \sum_{i=1}^{B} g_Z[i,j]\,(Y[i,j] - \mu[j])\left(-\frac{1}{2}\right)\mathrm{std}[j]^{-3},$$

$$\begin{aligned}
\frac{\partial \ell}{\partial \mu[j]} &= \sum_{i=1}^{B} \frac{\partial \ell}{\partial \hat{Y}[i,j]} \frac{\partial \hat{Y}[i,j]}{\partial \mu[j]} + \frac{\partial \ell}{\partial v[j]} \frac{\partial v[j]}{\partial \mu[j]} \\
&= \sum_{i=1}^{B} g_Z[i,j]\left(-\mathrm{std}[j]^{-1}\right) + \frac{\partial \ell}{\partial v[j]} \cdot \frac{-2}{B} \sum_{i=1}^{B} (Y[i,j] - \mu[j]),
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \ell}{\partial Y[i,j]} &= \frac{\partial \ell}{\partial \hat{Y}[i,j]} \frac{\partial \hat{Y}[i,j]}{\partial Y[i,j]} + \frac{\partial \ell}{\partial v[j]} \frac{\partial v[j]}{\partial Y[i,j]} + \frac{\partial \ell}{\partial \mu[j]} \frac{\partial \mu[j]}{\partial Y[i,j]} \\
&= g_Z[i,j]\,\mathrm{std}[j]^{-1} + \frac{\partial \ell}{\partial v[j]} \cdot \frac{2}{B}(Y[i,j] - \mu[j]) + \frac{\partial \ell}{\partial \mu[j]} \cdot \frac{1}{B}.
\end{aligned}$$

Key Takeaways (Q2)

- ε provides numerical stability by preventing division by tiny variances.
- Ignoring ε, BN makes each feature have mean β and variance γ^2.
- Backprop: ReLU mask $\Rightarrow$ BN param grads (β, γ) $\Rightarrow$ BN input grads .

Setup (notation)

- $h_i = \sigma(W_i h_{i-1} + b_i)$ for $i = 1, \ldots, L$; $h_0 = x$.
- Softmax readout: $y_k = \frac{e^{h_L[k]}}{\sum_j e^{h_L[j]}}$, CE loss: $\ell(\bar{y}, y) = -\sum_k \bar{y}[k] \log y[k]$.
- Shapes: $W_i \in \mathbb{R}^{D_i \times D_{i-1}}$, $b_i \in \mathbb{R}^{D_i \times 1}$, $h_i \in \mathbb{R}^{D_i \times 1}$.

3.1 : $\partial \ell / \partial h_L$ (softmax + CE)

Chain rule: $\dfrac{\partial \ell}{\partial h_L} = \dfrac{\partial \ell}{\partial y} \dfrac{\partial y}{\partial h_L}.$

- $\left[\dfrac{\partial \ell}{\partial y}\right]_k = -\dfrac{\bar{y}_k}{y_k}$ (one-hot $\bar{y}$).
- $\dfrac{\partial y}{\partial h_L} = \operatorname{diag}(y) - y\, y^\top.$

Result: $\dfrac{\partial \ell}{\partial h_L} = y - \bar{y}.$

3.1: Why $\dfrac{\partial y}{\partial h_L} = \operatorname{diag}(y) - yy^\top$?

With $y_k = \dfrac{e^{h_k}}{S}$, $S = \sum_{j=1}^{K} e^{h_j}$,

$$\frac{\partial y_k}{\partial h_i} = \frac{e^{h_k}\delta_{ki}}{S} - \frac{e^{h_k}}{S^2}\frac{\partial S}{\partial h_i} = y_k\delta_{ki} - y_k y_i.$$

Thus componentwise $\dfrac{\partial y_k}{\partial h_i} = y_k(\delta_{ki} - y_i)$, which stacks to

$$\frac{\partial y}{\partial h} = \operatorname{diag}(y) - yy^\top.$$

3.1: Gradient w.r.t. a hidden layer h_i (chain rule)

Layer relations: $z_{i+1} = W_{i+1} h_i + b_{i+1}, \; h_{i+1} = \sigma(z_{i+1})$.

Jacobian (reference form):

$$J_i \;:=\; \frac{\partial h_{i+1}}{\partial h_i} \;=\; \mathrm{diag}\big(\sigma'(z_{i+1})\big)\, W_{i+1} \;\in\; \mathbb{R}^{D_{i+1}\times D_i}.$$

Chain rule for hidden layers:

$$\frac{\partial \ell}{\partial h_i} \;=\; J_i^\top \frac{\partial \ell}{\partial h_{i+1}} \;=\; J_i^\top J_{i+1}^\top \cdots J_{L-1}^\top \frac{\partial \ell}{\partial h_L}$$

Using $\dfrac{\partial \ell}{\partial h_L} = y - \bar{y}$ from the previous slide,

$$\frac{\partial \ell}{\partial h_i} \;=\; W_{i+1}^\top \,\mathrm{diag}\big(\sigma'(z_{i+1})\big) \cdots W_L^\top \,\mathrm{diag}\big(\sigma'(z_L)\big)\,(y - \bar{y}).$$

3.2 Gradients w.r.t. parameters

Let $z_i = W_i h_{i-1} + b_i$ and $\delta_i := \partial \ell / \partial z_i$.

$$\delta_i = \sigma_i' \odot \frac{\partial \ell}{\partial h_i} \qquad (\mathbb{R}^{D_i \times 1})$$

$$\frac{\partial \ell}{\partial W_i} = \delta_i \, h_{i-1}^\top \ \in \ \mathbb{R}^{D_i \times D_{i-1}},$$

$$\frac{\partial \ell}{\partial b_i} = \delta_i \ \in \ \mathbb{R}^{D_i \times 1}.$$

3.3 Goal: Preserve $\mathrm{Var}[h_i]$

Objective. Choose the weight variance so that activation variance is stable across layers:

$$\mathrm{Var}[h_i] \;\approx\; \mathrm{Var}[h_{i-1}] \quad \text{for all } i.$$

We assume:

- $z_i = W_i h_{i-1} + b_i$, with $b_i = 0$ at init and W_i i.i.d., zero mean.
- Pre-activations z_i are approximately zero-mean and symmetric; activation $h_i = \mathrm{ReLU}(z_i)$.
- Fan-in $n = D_{i-1}$.

3.3 Deriving $\mathrm{Var}[z_i]$

$$z_i[j] \;=\; \sum_{k=1}^{D_{i-1}} w_{jk}\, h_{i-1}[k]$$

Using independence and zero-mean weights,

$$\begin{aligned}
\mathrm{Var}[z_i] &= D_{i-1}\ \mathrm{Var}[w_i h_{i-1}] \\
&= D_{i-1}\Big(\mathrm{Var}[w_i]\,\mathrm{Var}[h_{i-1}] + \mathrm{Var}[w_i]\big(\mathbb{E}[h_{i-1}]\big)^2 + \mathrm{Var}[h_{i-1}]\big(\mathbb{E}[w_i]\big)^2\Big) \\
&= D_{i-1}\Big(\mathrm{Var}[w_i]\,\mathrm{Var}[h_{i-1}] + \mathrm{Var}[w_i]\big(\mathbb{E}[h_{i-1}]\big)^2\Big) \qquad (\mathbb{E}[w_i] = 0) \\
&= \ D_{i-1}\ \mathrm{Var}[w_i]\ \mathbb{E}[h_{i-1}^2]\,.
\end{aligned}$$

3.3 $\mathbb{E}[z_i^2]$ via symmetry through ReLU

If w_{i-1} is symmetric about 0 and $b_{i-1} = 0$, then z_{i-1} is symmetric about 0. Hence for $h_i = \mathrm{ReLU}(z_{i-1})$,

$$\begin{aligned}\mathbb{E}[h_i^2] = \mathbb{E}\big[\mathrm{ReLU}^2(z_{i-1})\big] &= \int_{-\infty}^{+\infty} [\max(0, z_{i-1})]^2\, p(z_{i-1})\, dz_{i-1} \\ &= \int_0^{+\infty} z_{i-1}^2\, p(z_{i-1})\, dz_{i-1} = \tfrac{1}{2} \int_{-\infty}^{+\infty} z_{i-1}^2\, p(z_{i-1})\, dz_{i-1} \\ &= \tfrac{1}{2}\, \mathbb{E}[z_{i-1}^2] \;=\; \tfrac{1}{2}\, \mathrm{Var}[z_{i-1}] \quad (\mathbb{E}[z_{i-1}] = 0).\end{aligned}$$

So $q_i := \mathbb{E}[h_i^2] = \frac{1}{2}\, \mathrm{Var}[z_{i-1}]$.

3.3 Solve Var[w] with variance preservation

Now we have

$$\mathrm{Var}[h_i] = \left(\frac{D_{i-1}\,\sigma_w^2}{2}\right) \mathrm{Var}[h_{i-1}].$$

Enforcing $\mathrm{Var}[h_i] \approx \mathrm{Var}[h_{i-1}]$ yields

$$\sigma_w^2 = \frac{2}{D_{i-1}} = \frac{2}{\text{fan-in}}$$

— the standard **He/Kaiming** initialization for ReLU.

Recap

- Softmax gradient: $J = \text{diag}(y) - yy^\top \Rightarrow \partial\ell/\partial h_L = y - \bar{y}$.
- Parameter grads: $\partial\ell/\partial W_i = \delta_i h_{i-1}^\top, \quad \partial\ell/\partial b_i = \delta_i$.
- Reference-aligned steps: $\text{Var}[z_i] = D_{i-1}\,\text{Var}[w_i]\mathbb{E}[h_{i-1}^2]$ and $\mathbb{E}[z_i^2] = \frac{1}{2}\,\text{Var}[z_{i-1}]$.
- He init (ReLU): $\text{Var}[w] = 2/\text{fan-in}$.