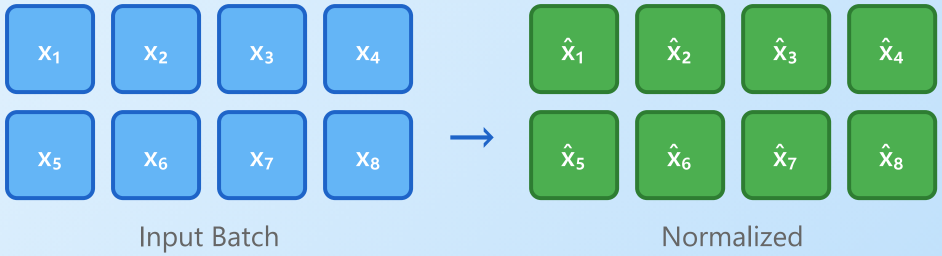
Batch Normalization (Batch Norm)

Normalizes Across Batch Dimension



Batch Norm Formula

$$\hat{x} = (x - \mu_{batch}) / \sqrt{(\sigma^2_{batch} + \varepsilon)}$$

$$y = \gamma \cdot \hat{x} + \beta$$

Learnable: γ (scale) & β (shift)

Key Benefits

- **Reduces** internal covariate shift significantly
- Enables **higher learning rates** and faster convergence
- Acts as **regularizer**, reduces need for dropout
- Normalizes inputs across the **batch dimension**
- **Learnable parameters** (γ, β) maintain expressiveness

✓ Most Widely Used

Standard normalization technique in **Convolutional Neural Networks (CNNs)**

🔬 How Batch Normalization Works

⚠️ Internal Covariate Shift Problem

As neural networks get deeper, the **data distribution changes** progressively through each layer, causing problems.

Sigmoid Activation Function

Steepest Region:
-2 to 2

Early Layers ✅

Data in -2~2 range

Fast Learning

Deep Layers ❌

Data drifts away

Slow Learning (Vanishing Gradient)

✨ Batch Normalization Solution

Re-normalize data at each layer to maintain optimal distribution throughout the network.

Applied at Every Layer:

1. Calculate batch mean and variance
2. Normalize data to -2~2 range
3. Scale with γ and shift with β
4. Pass to next layer

Without Normalization

🐌 Slow Convergence

With Normalization

🚀 Fast Convergence

💡 Key Insight 1: Alleviates Vanishing Gradient

Keeps data in the steepest region of sigmoid (-2~2), significantly reducing the vanishing gradient problem.

💡 Key Insight 2: Faster Training

Transforms the loss function from a valley shape to a bowl shape, finding the optimum much faster.

💡 Key Insight 3: Regularization Effect

Normalizing with slightly different mean/variance per batch naturally adds noise, providing regularization similar to Dropout.

📊 Numerical Calculation Example

1 Input Batch Data

Batch Size: **4 samples**

Input values: x = [1, 3, 5, 7]

2 Calculate Mean

$\mu = (1 + 3 + 5 + 7) / 4$

$\mu\_batch = 4.0$

3 Calculate Variance

$\sigma^2 = [(1-4)^2 + (3-4)^2 + (5-4)^2 + (7-4)^2] / 4$

$\sigma^2 = [9 + 1 + 1 + 9] / 4$

$\sigma^2\_batch = 5.0$

4 Normalize

$\hat{x} = (x - \mu) / \sqrt{(\sigma^2 + \varepsilon)}$

$\varepsilon = 0.001$ (for stability)

$\hat{x}_1 = (1 - 4) / \sqrt{5.001} = -1.34$

$\hat{x}_2 = (3 - 4) / \sqrt{5.001} = -0.45$

$\hat{x}_3 = (5 - 4) / \sqrt{5.001} = 0.45$

$\hat{x}_4 = (7 - 4) / \sqrt{5.001} = 1.34$

5 Apply Scale & Shift

Learnable Parameters:

γ (scale) = 2.0

Final Formula:

$y = \gamma \cdot \hat{x} + \beta$

β (shift) = 1.0

$y = 2.0 \cdot \hat{x} + 1.0$

✨ Final Output Values

y_1	y_2	y_3	y_4
-1.68	**0.10**	**1.90**	**3.68**

Original: [1, 3, 5, 7] → After Normalization: [-1.68, 0.10, 1.90, 3.68]