COS 514: Fundamentals of Deep Learning

Fall 2025

Instructor: Prof. Sanjeev Arora
TA: Gon Buzaglo

Assignment 3

Instructions:

- Submission deadline is November 3.
- You may collaborate in groups of up to **3** students.
- If you collaborate on a problem, you must clearly state the names of your collaborators at the beginning of the solution to that problem.
- All group members must declare that they contributed equally to the solutions.
- You must write up your own solutions independently in LaTeX. **Handwritten or scanned solutions will not be accepted.**
- Cite any resources (papers, textbooks, websites) that you use.
- Submit your assignment as a single PDF on gradescope.

Problems

Problem 3: Diffusion Models

Let $\{x_t\}_{t\geq 0} \subset \mathbb{R}^d$ follow

$$q(x_t \mid x_{t-1}) = \mathcal{N}\big(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t I\big), \qquad \beta_t \in (0,1),$$

and define $\alpha_t := 1 - \beta_t$, $\bar{\alpha}_t := \prod_{i=1}^{t} \alpha_i$ (with $\bar{\alpha}_0 := 1$).

(a) **Forward marginal.** Show that

$$q(x_t \mid x_0) = \mathcal{N}\big(x_t; \sqrt{\bar{\alpha}_t}\, x_0,\ (1 - \bar{\alpha}_t)I\big). \tag{1}$$

Hint: Linear combination of independent samples from two Gaussians is itself distributed like a Gaussian.

(b) **Reverse posterior and noise prediction.** This question studies in more detail why learning to predict the error from the noised image x_t in Diffusion models suffices to construct the reverse-step mean.

(i) Using Bayes' rule and (1), show

$$q(x_{t-1} \mid x_t, x_0) \propto \exp\Big[-\tfrac{1}{2}\Big(\frac{\|x_t - \sqrt{\alpha_t} x_{t-1}\|^2}{\beta_t} + \frac{\|x_{t-1} - \sqrt{\bar{\alpha}_{t-1}} x_0\|^2}{1 - \bar{\alpha}_{t-1}}\Big)\Big].$$

(ii) Complete the square in x_{t-1} to derive the mean

$$\tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t}\, x_0 + \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}\, x_t.$$

(iii) From (1), write $x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$, so $x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\big(x_t - \sqrt{1 - \bar{\alpha}_t}\, \epsilon\big)$. Substitute this into $\tilde{\mu}_t$ to express $\tilde{\mu}_t = \tilde{\mu}_t(x_t, \epsilon)$.

Problem 2: Orthogonal Equivariance

A learning algorithm A is *orthogonally equivariant* if for any orthogonal R, training A on inputs Rx yields predictions R times those from training on x. Prove that both SGD and SGD with momentum are orthogonally equivariant. (*Hint:* Track how gradients transform under $x \mapsto Rx$.)

Problem 3: ReLU and Vanishing Gradients

In a deep network $\hat{y} = f_L(\cdots f_1(x))$, gradients are products of Jacobians. Compare gradient propagation in networks using ReLU vs. sigmoid:

(a) Why does ReLU reduce—but not fully eliminate—the vanishing gradient problem?

(b) Discuss roles of activation derivatives, initialization, and depth.

(c) Give one training failure mode specific to ReLU and how it's mitigated.

Problem 4: Batch Normalization

(a) If BN is used without affine parameters (γ, β), does the network lose expressive power? Can other layers replicate their effect?

(b) In a block ReLU $\to$ Linear $\to$ BN, scaling pre-ReLU activations by c multiplies u by c. For $c > 0$ vs. $c < 0$, how does this affect BN's scale invariance? What role does the bias b play?

Problem 5: ResNets and Normalization

Let

$$H(x) = x + F(x), \quad F(x) = \text{Conv}_2(\text{ReLU}(\text{BN}_1(\text{Conv}_1(x)))).$$

(a) Does scaling $W_1 \mapsto cW_1$ change $H(x)$? Explain how scale invariance of BN interacts with ReLU, Conv_2, and the residual sum.

(b) Consider $x_{l+1} = \frac{1}{\sqrt{2}}(x_l + F(x_l))$. Assuming $\text{Var}(F(x_l)) \approx \text{Var}(x_l)$ and they're uncorrelated, does this fix variance explosion? What drawback might it introduce compared to the standard ResNet block?