Aprendizado Profundo

Prova, Unidade 5

Prof. Lucas S. Kupssinskü

Data: _____

Nome: _____

Quadro de respostas (questões objetivas):

Questão	a	b	c	d	e	f	g	h
1								
2								
3								
4								
5								

Questões Objetivas (6.0 pts)

Cada questão possui apenas uma alternativa correta. Marque a letra escolhida no quadro de respostas.

1) (1.2) Considere um processo de difusão com $T = 2$ e *noise schedule* $\boldsymbol{\beta} = (0.2,\ 0.2)$. Para a instância unidimensional $x = 3$ e ruído $\epsilon = -1$, calcule o produto acumulado α_2 e o valor z_2 obtido ao aplicar o kernel de difusão que amostra diretamente o passo $t = 2$ a partir de x e ϵ. Quais são, respectivamente, os valores de α_2 e z_2?

   a) $\alpha_2 = 0.64$ e $z_2 = 1.56$.

   b) $\alpha_2 = 0.60$ e $z_2 \approx 1.69$.

   c) $\alpha_2 = 0.64$ e $z_2 = 1.80$.

   d) $\alpha_2 = 0.64$ e $z_2 = 3.00$.

   e) $\alpha_2 = 0.04$ e $z_2 \approx -0.38$.

   f) $\alpha_2 = 0.64$ e $z_2 = 1.00$.

   g) $\alpha_2 = 0.80$ e $z_2 \approx 2.24$.

   h) $\alpha_2 = 0.64$ e $z_2 = 1.40$.

2) (1.2) Sobre as asserções abaixo a respeito de *normalizing flows*:

   I. Em uma camada *affine coupling*, o log-determinante da Jacobiana vale $\sum_i s_i$, e a *coupling network* que produz $(\mathbf{s}, \mathbf{t})$ não precisa ser inversível, pois nunca é invertida, apenas reavaliada a partir de $\mathbf{x}^A$.

   II. Empilhar várias camadas de *linear flow* $f(\mathbf{x}) = \boldsymbol{\theta}\mathbf{x} + \mathbf{b}$ com matrizes $\boldsymbol{\theta}$ distintas aumenta progressivamente a expressividade do modelo, permitindo representar densidades que uma única camada linear não consegue representar.

   III. Por exigirem uma bijeção, *normalizing flows* preservam a dimensão ($\dim \mathbf{z} = \dim \mathbf{x}$), de modo que não há gargalo de compressão como no latente de um VAE.

   a) Apenas I está correta.

   b) Apenas II está correta.

   c) Apenas III está correta.

d) Apenas I e II estão corretas.

e) Apenas I e III estão corretas.

f) Apenas II e III estão corretas.

g) Todas estão corretas.

h) Nenhuma está correta.

3) (1.2) Você treina uma GAN com a *loss* original do gerador, $L[\boldsymbol{\theta}] = \sum_j \ln\big[1 - \sigma\big(f\big[g[z^{(j)}, \boldsymbol{\theta}], \boldsymbol{\phi}\big]\big)\big]$. Nas primeiras épocas, a acurácia do discriminador sobe para perto de 100% e $\sigma(f[x^*, \boldsymbol{\phi}]) \approx 0$ em todas as amostras geradas; a *loss* do gerador fica praticamente constante e $\|\nabla_{\boldsymbol{\theta}} L\| \approx 0$. Qual é o diagnóstico mais consistente com esses sintomas e a mitigação mais direta?

a) O discriminador venceu o jogo cedo demais; com $\sigma \approx 0$ a *loss* original satura e o gradiente do gerador some; trocar para a *non-saturating loss* $-\ln\sigma(\cdot)$ recupera o sinal.

b) O gerador entrou em *mode collapse* e produz amostras idênticas entre si; adicionar *minibatch discrimination* restaura a diversidade das amostras e o gradiente.

c) A taxa de aprendizado do gerador está alta demais e ele saltou por cima do mínimo; reduzi-la em uma ordem de grandeza devolve o progresso normal do treinamento.

d) O discriminador sofre de *overfitting* sobre as imagens reais do conjunto de treino; aplicar *Dropout* nas camadas do gerador equilibra o jogo adversarial.

e) O treinamento atingiu o equilíbrio de Nash ($P^* = P$), no qual o gradiente nulo é o comportamento esperado; basta interromper o treino e amostrar do gerador.

f) Os *manifolds* de P e P^* estão sobrepostos demais desde o início; reduzir a dimensão do latente z separa as distribuições e recria o sinal de gradiente.

g) A última camada do gerador deveria usar ReLU em vez de *Tanh*; o intervalo de saída incorreto impede o discriminador de aprender qualquer fronteira útil.

h) O *batch size* está pequeno demais para estimar a esperança da *loss*; aumentá-lo reduz a variância do gradiente e elimina a saturação observada.

4) (1.2) Em um modelo difusor, a U-Net que estima o ruído recebe como entrada não apenas o latente z_t, mas também um *time embedding* construído a partir do passo t. Por que essa informação de t é necessária?

a) Porque o *embedding* de t funciona como *positional encoding* espacial, informando à U-Net a posição de cada *pixel* dentro da imagem ruidosa.

b) Porque o *embedding* de t substitui o *noise schedule* β_t, dispensando a escolha de hiperparâmetros do processo de difusão durante o treinamento.

c) Porque o *embedding* de t injeta a estocasticidade necessária na rede, fazendo o papel do termo de ruído $\sigma_t\,\epsilon$ usado na amostragem reversa.

d) Porque t indexa qual das T redes completamente independentes deve ser ativada, e o *embedding* seleciona o subconjunto de pesos correspondente ao passo.

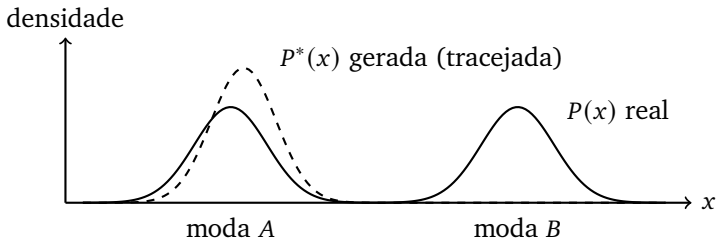
e) Porque o *embedding* de t atua como um regularizador da U-Net, prevenindo que ela memorize o conjunto de treinamento ao longo dos T passos.

f) Porque a magnitude do ruído em z_t varia com t conforme o *schedule*; a rede precisa saber o nível de ruído para calibrar a remoção em cada passo.

g) Porque a propriedade markoviana do processo de difusão só se mantém se o *decoder* conhecer t; sem o *embedding* a cadeia deixa de ser Markov.

h) Porque sem o *embedding* de t a *loss* MSE $|g_t[z_t, \boldsymbol{\theta}_t] - \epsilon|^2$ deixa de ser diferenciável em relação aos parâmetros da U-Net.

5) (1.2) A figura abaixo mostra a densidade dos dados reais $P(x)$ (linha cheia, com duas modas A e B) e a densidade gerada $P^*(x)$ (linha tracejada) de uma GAN durante o treinamento.



Considerando a decomposição $D_{JS}[P^*\|P] = \underbrace{\frac{1}{2}D_{KL}[P^*\|M]}_{\text{qualidade}} + \underbrace{\frac{1}{2}D_{KL}[P\|M]}_{\text{cobertura}}$, com $M = \frac{1}{2}(P + P^*)$, qual afirmação descreve corretamente o fenômeno ilustrado e o comportamento do gradiente do gerador?

a) Trata-se de *vanishing gradient* por suportes disjuntos: como P^* não toca a moda B, D_{JS} já atingiu o teto $\ln 2$ e nenhum dos dois termos produz gradiente para o gerador.

b) Trata-se de *mode collapse*: na moda B, onde $P^* \approx 0$, o integrando da cobertura satura em $\frac{1}{2}P \ln 2$, quase insensível a P^*, então o incentivo para cobrir B é fraco.

c) Trata-se de *mode collapse*: o termo de qualidade é o que satura na moda B, e o gerador recebe incentivo forte para cobrir B, mas o discriminador ótimo bloqueia esse movimento.

d) Trata-se de *posterior collapse*: a KL entre o posterior aproximado e o *prior* foi a zero, então o latente deixou de carregar informação sobre qual das duas modas gerar.

e) Trata-se de *overfitting* do discriminador: P^* concentrada na moda A indica que o discriminador memorizou essa moda; reiniciar os pesos ϕ recupera a moda B.

f) Trata-se de *mode collapse*: a KL da cobertura diverge para infinito na moda B, o gradiente correspondente domina o treino e força o gerador a cobrir B nas iterações seguintes.

g) Trata-se de *vanishing gradient*: a sigmóide do discriminador satura na moda A, onde P e P^* se sobrepõem, e o treinamento congela com $D_{JS} = 0$.

h) Trata-se de *mode collapse*: como D_{JS} é simétrica, qualidade e cobertura geram gradientes de mesma magnitude, e o colapso resulta apenas da inicialização infeliz do gerador.

Questões Dissertativas (4.0 pts)

6) (2.0) Marque **Verdadeiro** ou **Falso** nas asserções abaixo. Em caso **Falso**, proponha uma modificação *não-trivial* que torna a asserção verdadeira (a simples negação **não** é uma modificação válida). Cada item vale 0.4 pts.

a) ( ) V ( ) F No *WGAN*, a restrição de 1-Lipschitz do crítico é imposta na prática por *gradient clipping*: após cada atualização, os gradientes do crítico são truncados para o intervalo $[-c, c]$.

b) ( ) V ( ) F Em uma camada *affine coupling*, a *coupling network* que produz os vetores **s** e **t** precisa ser inversível, caso contrário a camada como um todo deixa de ser inversível e o *flow* não pode ser avaliado.

c) ( ) V ( ) F O *reparametrization trick* reescreve a amostragem do latente como $z^* = \mu + \Sigma^{1/2}\epsilon^*$, com $\epsilon^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, de modo que a estocasticidade fica concentrada em ϵ^* e o caminho por μ e $\Sigma^{1/2}$ permanece diferenciável em relação a ϕ.

d) ( ) V ( ) F No modelo difusor, o *encoder* que adiciona ruído possui parâmetros treináveis ajustados em conjunto com o *decoder*, de forma análoga ao *encoder* de um VAE.

e) ( ) V ( ) F No trilema dos modelos geradores, os modelos difusores ocupam o canto *qualidade + cobertura*, pagando com amostragem lenta (tipicamente $T \approx 1000$ passos pela U-Net).

7) (2.0) Considere um VAE com latente bidimensional ($D_z = 2$) e *decoder* $p(x \mid z, \boldsymbol{\theta}) = \mathcal{N}_x(f[z, \boldsymbol{\theta}], \mathbf{I})$. Para a entrada $x = (2.4,\ -1.0)$, o *encoder* $g[x, \phi]$ produz

$$\mu = (2,\ -1), \qquad \Sigma = \mathrm{diag}(0.25,\ 4) \quad (\text{logo } \sigma = (0.5,\ 2)).$$

Use a forma fechada $D_{KL}[\mathcal{N}(\mu, \Sigma) \,\|\, \mathcal{N}(\mathbf{0}, \mathbf{I})] = \frac{1}{2}(\mathrm{Tr}[\Sigma] + \mu^\top \mu - D_z - \log\det\Sigma)$.

a) (0.5) Calcule $D_{KL}[q(z \mid x, \phi) \,\|\, p(z)]$, mostrando o valor de cada um dos quatro termos.

b) (0.5) Foi amostrado $\epsilon^* = (0.2,\ -0.3)$. Calcule z^* pelo *reparametrization trick*.

c) (0.5) O *decoder* devolve $f[z^*, \boldsymbol{\theta}] = (2.0,\ -1.3)$. Calcule o termo de reconstrução $\frac{1}{2}\|x - f[z^*, \boldsymbol{\theta}]\|^2$ e monte o valor aproximado de $-$ELBO $\approx$ reconstrução + D_{KL}.

d) (0.5) Suponha que, para todas as entradas, o *encoder* passasse a produzir $\Sigma = \mathrm{diag}(10^{-6},\ 10^{-6})$. O que acontece com o termo $-\log\det\Sigma$ da KL e qual comportamento indesejado do *encoder* essa penalização evita?