Instruções: Responda às questões de múltipla escolha assinalando apenas uma alternativa. Nas dissertativas, apresente o raciocínio passo a passo, incluindo fórmulas e cálculos intermediários quando solicitado. Mantenha a notação dos slides: $g[z, \theta]$ para o gerador (com latente z e parâmetros θ), $f[x, \phi]$ para o discriminador/crítico, $\sigma(\cdot)$ para a sigmóide, $P(x)$ para a distribuição real, $P^*(x)$ para a distribuição gerada, $M(x) = \frac{1}{2}(P + P^*)$ para a mistura, $D^*(x) = \frac{P(x)}{P(x)+P^*(x)}$ para o discriminador ótimo, D_{JS} para a divergência de Jensen–Shannon e D_w para a distância de Wasserstein. Nas questões sobre Wasserstein, as distribuições comparadas são denotadas P e Q, e o plano de transporte é T, como nos slides. Questões marcadas com ☠ são consideradas difíceis.

Questões de Múltipla Escolha.

I) Considere a categoria a, com $P(a) = 0.40$ e $P^*(a) = 0.10$. Qual é o valor do discriminador ótimo $D^*(a)$ nesse ponto?

(a) 0.80.

(b) 0.20.

(c) 0.40.

(d) 0.50.

II) Para uma única amostra **real** ($y = 1$), o discriminador devolve $\sigma(f[x, \phi]) = 0.20$. Qual é a contribuição dessa amostra à *loss* de entropia cruzada binária do discriminador?

(a) 0.200.

(b) ≈ 0.223.

(c) ≈ 1.609.

(d) 0.800.

III) Seja $d = f[g[z, \theta], \phi]$ o logit do discriminador na amostra gerada. No ponto $d = -3$, quais são, respectivamente, as magnitudes dos gradientes da *loss* original do gerador $L_{\text{orig}} = \ln[1 - \sigma(d)]$ e da *loss non-saturating* $L_{\text{NS}} = -\ln[\sigma(d)]$ em relação a d? Use $\sigma(-3) \approx 0.047$.

(a) 0.953 e 0.047.

(b) 0.047 e 0.953.

(c) 0.500 e 0.500.

(d) 0.047 e 0.047.

IV) Sobre o suporte $\{a, b\}$, considere $P = (0.5, 0.5)$ e $P^* = (0.0, 1.0)$. Qual é o valor de $D_{JS}(P^* \| P)$ em nats? Use $\ln 2 \approx 0.693$, $\ln(4/3) \approx 0.288$ e $\ln(2/3) \approx -0.405$.

(a) ≈ 0.14.

(b) ≈ 0.43.

(c) ≈ 0.69.

(d) ≈ 0.22.

V) Sobre o suporte $\{0, 1, 2\}$, considere $P = (0.5, 0.3, 0.2)$ e $Q = (0.2, 0.3, 0.5)$. Usando o custo $|i - j|$, qual é o valor de $D_w(P \| Q)$ obtido pelo plano de transporte ótimo?

(a) 0.0.

(b) 1.2.

(c) 0.3.

(d) 0.6.

VI) Você implementa uma DCGAN cujas imagens reais estão normalizadas em $[-1, 1]$. Por engano, troca a ativação *Tanh* da última camada do gerador por *ReLU*, mantendo o resto da arquitetura. Qual o efeito mais provável no treinamento?

(a) ReLU produz amostras de qualidade semelhante à *Tanh* desde que haja *BatchNorm* imediatamente antes da última convolução.

(b) O gerador fica restrito a $[0, \infty)$ e nunca produz os valores negativos das imagens reais, e o treinamento degenera.

(c) ReLU acelera o treinamento por eliminar o *vanishing gradient* típico da *Tanh* em saturação.

(d) Como o gerador da DCGAN não usa camadas *fully connected*, a ativação da última convolução não influencia o intervalo da saída.

VII) Em *WGAN-GP*, o termo de penalidade adicionado à *loss* é $\lambda\,(\|\nabla_{\hat{x}} f[\hat{x}, \phi]\|_2 - 1)^2$ avaliado em um ponto interpolado $\hat{x}$. Para $\lambda = 10$ e $\|\nabla_{\hat{x}} f[\hat{x}, \phi]\|_2 = 3$, qual é a contribuição desse ponto à *loss*?

(a) 4.

(b) 90.

(c) 40.

(d) 20.

VIII) ☠️ Sobre o suporte $\{a, b, c, d\}$, considere $P = (0.5,\ 0.5,\ 0,\ 0)$ e $P^* = (0,\ 0,\ 0.4,\ 0.6)$. Qual é o valor de $D_{JS}(P^* \| P)$ em nats? Use a convenção $0 \ln 0 = 0$.

(a) ≈ 0.35.

(b) ≈ 1.39.

(c) 0.

(d) ≈ 0.69.

IX) Em uma equipe de pesquisa, alguém descreve um modelo gerador com a seguinte combinação de propriedades: (i) gera amostras nítidas em uma única passagem pela rede, (ii) frequentemente ignora modas inteiras do conjunto real durante o treinamento, (iii) o ajuste de hiperparâmetros é notoriamente instável. A qual família o modelo descrito corresponde, dentro do trilema apresentado nos slides?

(a) GAN.

(b) Modelo difusor (*diffusion model*).

(c) VAE.

(d) *Normalizing flow*.

X) ☠️ Considere $P = (0.6, 0.3, 0.1)$ e $Q = (0.1, 0.3, 0.6)$ sobre o suporte $\{0, 1, 2\}$, cuja distância de Wasserstein é $D_w = 1$. Qual dos candidatos abaixo para o crítico $f = (f_0, f_1, f_2)$ satisfaz a restrição 1-Lipschitz $|f_{i+1} - f_i| \leq 1$ *e* atinge D_w no dual $\sum_i P(i)\, f_i - \sum_j Q(j)\, f_j$?

(a) $f = (0, 1, 2)$.

(b) $f = (2, 1, 0)$.

(c) $f = (10, 5, 0)$.

(d) $f = (1, 1, 1)$.

Gabarito - Objetivas

I) a

II) c

III) b

IV) d

V) d

VI) b

VII) c

VIII) d

IX) a

X) b

Resoluções - Objetivas

I) $D^*(a) = \frac{P(a)}{P(a)+P^*(a)} = \frac{0.40}{0.40+0.10} = \frac{0.40}{0.50} = 0.80$. A (b) corresponde a inverter os papéis no numerador, $\frac{P^*}{P+P^*} = 0.20$. A (c) corresponde a devolver $P(a)$ sem normalizar pela soma. A (d) corresponde ao palpite *chance* de 0.5, ignorando os dados.

II) Como o rótulo é $y = 1$ (real), a contribuição é $-\ln[\sigma(f[x, \phi])] = -\ln(0.20) \approx 1.609$ (alternativa c). A (a) reporta σ diretamente, sem aplicar o logaritmo. A (b) usa $-\ln(1 - \sigma) = -\ln(0.80) \approx 0.223$, que seria correto se a amostra fosse gerada ($y = 0$). A (d) reporta $1 - \sigma$, também sem logaritmo.

III) Diferenciando termo a termo: $\partial L_{\text{orig}}/\partial d = -\sigma(d)$, com módulo $\sigma(d) = 0.047$. Para a *non-saturating*, $\partial L_{\text{NS}}/\partial d = -(1 - \sigma(d))$, com módulo $1 - 0.047 = 0.953$ (alternativa b). A (a) troca os dois rótulos. A (c) assume gradiente constante de 0.5, que vale apenas no equilíbrio $\sigma = \frac{1}{2}$. A (d) atribui o mesmo valor a ambas, ignorando que a NS foi construída exatamente para não saturar quando $\sigma \to 0$.

IV) Com $M = (0.25, 0.75)$: $D_{KL}(P\|M) = 0.5 \ln 2 + 0.5 \ln(2/3) \approx 0.347 - 0.203 = 0.144$, e $D_{KL}(P^*\|M) = \ln(4/3) \approx 0.288$ (o termo em a vale 0 pela convenção $0 \ln 0 = 0$). Logo $D_{JS} = \frac{1}{2}(0.144 + 0.288) \approx 0.22$ nats (alternativa d). A (b) corresponde a somar as duas KL sem o fator $\frac{1}{2}$ externo. A (c) corresponde a tratar as distribuições como suporte disjunto e usar o limite $\ln 2$. A (a) corresponde a usar apenas $D_{KL}(P\|M)$.

V) As diferenças $P - Q = (+0.3,\ 0,\ -0.3)$ indicam que 0.3 de massa deve sair do índice 0 e chegar ao índice 2. O plano ótimo concentra esse movimento em uma única rota, custo $0.3 \cdot |0 - 2| = 0.6$ (alternativa d). A (b) corresponde a contabilizar duas vezes o mesmo transporte (origem e destino). A (c) corresponde a usar distância 1 em vez de 2 entre os extremos. A (a) corresponde ao plano *diagonal* (T_{ij} não-nulo apenas em $i = j$), que viola as restrições marginais quando $P \neq Q$.

VI) Imagens reais ocupam $[-1, 1]$; um gerador com *ReLU* na saída só consegue produzir valores em $[0, \infty)$, e em particular nunca em $[-1, 0)$. O discriminador descobre essa pista determinística (qualquer pixel negativo $\Rightarrow$ real) e fica próximo de $\sigma \approx 0$ nas amostras geradas, fazendo o gradiente da *loss* original do gerador colapsar (alternativa b). A (a) ignora a incompatibilidade entre o intervalo de saída e o domínio dos dados. A (c) confunde *vanishing gradient* de saturação interna com o problema de *range* da saída, são fenômenos distintos. A (d) confunde a recomendação da DCGAN, ausência de *fully connected* ocultas, com indiferença em relação à ativação final (que determina o intervalo da saída).

VII) Penalidade $= \lambda\,(\|\nabla f\|_2 - 1)^2 = 10 \cdot (3 - 1)^2 = 10 \cdot 4 = 40$ (alternativa c). A (b) corresponde a $\lambda\|\nabla f\|^2 = 10 \cdot 9 = 90$, esquecendo o -1 dentro do quadrado. A (a) corresponde a $(3 - 1)^2 = 4$, esquecendo λ. A (d) corresponde a $\lambda(3 - 1) = 20$, esquecendo de elevar ao quadrado.

VIII) Com $M = (0.25,\ 0.25,\ 0.20,\ 0.30)$: nos pontos em que $P > 0$, $\frac{P}{M} = 2$; nos pontos em que $P^* > 0$, $\frac{P^*}{M} = 2$. Logo $D_{KL}(P\|M) = (0.5 + 0.5)\ln 2 = \ln 2$ e, analogamente, $D_{KL}(P^*\|M) = (0.4 + 0.6)\ln 2 = \ln 2$. Portanto $D_{JS} = \frac{1}{2}(\ln 2 + \ln 2) = \ln 2 \approx 0.69$ nats, o teto da Jensen–Shannon (alternativa d). A (b) corresponde a somar as duas KL sem o fator $\frac{1}{2}$. A (c) corresponde a confundir suporte disjunto com distribuições iguais. A (a) corresponde a aplicar o fator $\frac{1}{2}$ duas vezes.

IX) A combinação "rápido na amostragem + má cobertura + instável" é o canto *qualidade + velocidade* do trilema, ocupado por GANs. A (b) descreve difusores, que estão no canto *qualidade + cobertura* (custosos na amostragem). A (c) e a (d) descrevem famílias do canto *velocidade + cobertura*, que costumam ter cobertura boa e estabilidade alta, em troca de qualidade mais limitada.

X) Verificando cada candidato: $f = (2, 1, 0)$ (alternativa b) é 1-Lipschitz ($|1 - 2| = |0 - 1| = 1$) e produz $(2 \cdot 0.6 + 1 \cdot 0.3 + 0 \cdot 0.1) - (2 \cdot 0.1 + 1 \cdot 0.3 + 0 \cdot 0.6) = 1.5 - 0.5 = 1$, atingindo D_w. A (a) é 1-Lipschitz mas tem orientação invertida, $0.5 - 1.5 = -1$; o dual exige o máximo, não o mínimo. A (c) atinge soma 5, porém viola Lipschitz ($|5 - 10| = 5 > 1$), ficando fora da região viável. A (d) é constante: 1-Lipschitz, mas produz soma 0, não distinguindo P de Q.

Questões Dissertativas.

I) Partindo da *loss* normalizada do discriminador,

$$L[\phi] = \mathbb{E}_{x^*}[\ln[1 - \sigma(f[x^*, \phi])]] + \mathbb{E}_x[\ln[\sigma(f[x, \phi])]]\,,$$

mostre, em passos algébricos, que com o discriminador ótimo $D^*(x) = \frac{P(x)}{P(x)+P^*(x)}$ a *loss* se reduz a

$$L[\phi] = 2\,D_{JS}(P^*\|P) - 2\ln 2.$$

Indique explicitamente: (a) onde a hipótese de iguais prioris ($P(\text{real}) = P(\text{gen}) = \frac{1}{2}$) entra; (b) onde aparece o "truque do fator 2"; (c) por que a constante $-2\ln 2$ não afeta o arg max em relação a ϕ nem o arg min em relação a θ (lembre que, escrita sem o sinal de menos, essa é a forma que o discriminador maximiza e o gerador minimiza).

II) ☠ Seja $d = f[g[z, \theta], \phi]$ o logit do discriminador na amostra gerada. Defina

$$L_{\text{orig}}[\theta] = \ln[1 - \sigma(d)] \qquad \text{e} \qquad L_{\text{NS}}[\theta] = -\ln[\sigma(d)],$$

ambas a serem minimizadas pelo gerador.

(a) Derive, mostrando todos os passos, que $\partial L_{\text{orig}}/\partial d = -\sigma(d)$ e $\partial L_{\text{NS}}/\partial d = -(1 - \sigma(d))$. (Lembre que $\sigma'(d) = \sigma(d)(1 - \sigma(d))$.)

(b) Mostre que as duas formulações apontam o gradiente *na mesma direção* (ambas aumentam d) e que, no equilíbrio adversarial $\sigma(d) = \frac{1}{2}$, as magnitudes dos gradientes coincidem e valem $\frac{1}{2}$ em ambos os casos.

(c) Avalie ambas as magnitudes em $d \in \{-6,\ 0,\ +6\}$ e organize em uma pequena tabela. Comente, em uma frase, o caso $d \to -\infty$ (discriminador vencendo o jogo): qual das duas *losses* produz gradiente útil para treinar o gerador?

III) A divergência de Jensen–Shannon entre real e gerada decompõe-se como

$$D_{JS}(P^*\|P) = \underbrace{\tfrac{1}{2}D_{KL}\left[P^*\,\middle\|\,M\right]}_{\text{qualidade}} + \underbrace{\tfrac{1}{2}D_{KL}\left[P\,\middle\|\,M\right]}_{\text{cobertura}}, \quad M = \tfrac{1}{2}(P + P^*).$$

(a) Para $P = (0.5,\ 0.3,\ 0.2)$ e $P^* = (0.2,\ 0.5,\ 0.3)$ sobre $\{a, b, c\}$, calcule M, $D_{KL}(P\|M)$, $D_{KL}(P^*\|M)$ e D_{JS}. Use $\ln 2 \approx 0.693$ e apresente o resultado em nats com três casas decimais.

(b) Suponha que, em alguma região R, $P^*(x) \ll P(x)$ (o gerador ignora uma moda). Mostre que, nessa região, o integrando do termo de *cobertura* é aproximadamente $P(x)\ln 2$, ou seja, fica **limitado e pouco sensível** a pequenos aumentos de $P^*(x)$.

(c) Em duas ou três frases, conecte (b) ao fenômeno de *mode collapse*: por que o gerador prefere "ganhar qualidade" nas modas que já cobre a "ganhar cobertura" em modas que ignora?

Gabarito

I) Partimos de

$$L[\phi] = \mathbb{E}_{x^*}[\ln[1-\sigma(f[x^*,\phi])]] + \mathbb{E}_x[\ln[\sigma(f[x,\phi])]] = \int P^*(x)\ln[1-\sigma(f)]\,dx + \int P(x)\ln[\sigma(f)]\,dx.$$

O discriminador ótimo é obtido pela regra de Bayes assumindo $P(\text{real}) = P(\text{gen}) = \frac{1}{2}$ **(item (a))**, o que cancela os *priors* e produz $\sigma(f[x,\phi]) = \frac{P(x)}{P(x)+P^*(x)}$, com complemento $1-\sigma = \frac{P^*(x)}{P(x)+P^*(x)}$. Substituindo,

$$L = \int P^*(x)\ln\left[\frac{P^*(x)}{P(x)+P^*(x)}\right]dx + \int P(x)\ln\left[\frac{P(x)}{P(x)+P^*(x)}\right]dx.$$

Introduzimos a mistura $M(x) = \frac{1}{2}(P(x)+P^*(x))$, de modo que $P(x)+P^*(x) = 2M(x)$. Multiplicando e dividindo por 2 dentro de cada logaritmo **(item (b))**, cada razão se reescreve como $\frac{2P^*(x)}{P(x)+P^*(x)} = \frac{P^*(x)}{M(x)}$ (e analogamente para $P(x)$):

$$L = \int P^*(x)\left(\ln\left[\frac{P^*(x)}{M(x)}\right] - \ln 2\right)dx + \int P(x)\left(\ln\left[\frac{P(x)}{M(x)}\right] - \ln 2\right)dx.$$

Os termos $-\ln 2$ saem das integrais e somam $-2\ln 2$, pois $\int P^*(x)\,dx = \int P(x)\,dx = 1$. As integrais restantes são exatamente $D_{KL}(P^*\|M)$ e $D_{KL}(P\|M)$. Logo

$$L = D_{KL}(P^*\|M) + D_{KL}(P\|M) - 2\ln 2 = 2\,D_{JS}(P^*\|P) - 2\ln 2.$$

Item (c): $-2\ln 2$ é constante em ϕ e em θ. Como essa forma da *loss* (sem o sinal de menos) é a log-verossimilhança que o discriminador *maximiza* e que o gerador *minimiza* (minimizar L equivale a minimizar D_{JS}), a constante não altera $\arg\max_\phi$ nem $\arg\min_\theta$.

II) **(a)** Usando $\sigma'(d) = \sigma(d)(1-\sigma(d))$:

$$\frac{\partial L_{\text{orig}}}{\partial d} = \frac{\partial}{\partial d}\ln[1-\sigma(d)] = \frac{-\sigma'(d)}{1-\sigma(d)} = \frac{-\sigma(1-\sigma)}{1-\sigma} = -\sigma(d).$$

$$\frac{\partial L_{\text{NS}}}{\partial d} = -\frac{\partial}{\partial d}\ln[\sigma(d)] = -\frac{\sigma'(d)}{\sigma(d)} = -\frac{\sigma(1-\sigma)}{\sigma} = -(1-\sigma(d)).$$

(b) Ambas as derivadas são *negativas* para $\sigma(d) \in (0,1)$, portanto a descida de gradiente em d aumenta d em ambos os casos: as duas *losses* empurram o gerador na mesma direção, mudando apenas a intensidade. No equilíbrio adversarial $\sigma(d) = \frac{1}{2}$ (ponto em que o discriminador não distingue real de gerado), as duas magnitudes coincidem: $|\partial L_{\text{orig}}/\partial d| = \sigma(d) = \frac{1}{2}$ e $|\partial L_{\text{NS}}/\partial d| = 1-\sigma(d) = \frac{1}{2}$.

(c)

d	$\sigma(d)$	$\lvert\partial L_{\text{orig}}/\partial d\rvert = \sigma(d)$	$\lvert\partial L_{\text{NS}}/\partial d\rvert = 1-\sigma(d)$
-6	≈ 0.0025	≈ 0.0025	≈ 0.9975
0	0.5000	0.5000	0.5000
$+6$	≈ 0.9975	≈ 0.9975	≈ 0.0025

Quando $d \to -\infty$ (discriminador vencendo, $\sigma \to 0$), $|\partial L_{\text{orig}}/\partial d| \to 0$ e a *loss* original satura, ao passo que $|\partial L_{\text{NS}}/\partial d| \to 1$ e o gerador continua recebendo gradiente útil; por isso a NS é o *default* na prática.

III) **(a)** $M = \frac{1}{2}(P + P^*) = (0.35,\ 0.40,\ 0.25)$.

$$D_{KL}(P\|M) = 0.5 \ln \tfrac{0.5}{0.35} + 0.3 \ln \tfrac{0.3}{0.40} + 0.2 \ln \tfrac{0.2}{0.25} \approx 0.178 - 0.086 - 0.045 = 0.047.$$

$$D_{KL}(P^*\|M) = 0.2 \ln \tfrac{0.2}{0.35} + 0.5 \ln \tfrac{0.5}{0.40} + 0.3 \ln \tfrac{0.3}{0.25} \approx -0.112 + 0.112 + 0.055 = 0.054.$$

$$D_{JS}(P^*\|P) = \tfrac{1}{2}(0.047 + 0.054) \approx 0.051 \text{ nats.}$$

(b) Em região R onde $P^*(x) \ll P(x)$, temos $M(x) = \frac{P(x)+P^*(x)}{2} \approx \frac{P(x)}{2}$. Logo $\frac{P(x)}{M(x)} \approx 2$, e o integrando da cobertura fica

$$P(x) \ln \tfrac{P(x)}{M(x)} \approx P(x) \ln 2,$$

limitado (cresce apenas com $P(x)$, não com a discrepância em si) e **pouco sensível** a pequenas perturbações em $P^*(x)$ (a derivada em $P^*(x)$ envolve $\frac{1}{P(x)+P^*(x)}$, que é pequena quando $P(x)$ domina).

(c) O termo de qualidade ($D_{KL}(P^*\|M)$) explode se o gerador colocar massa em pontos onde P é baixa: o gradiente pune **fortemente** amostras implausíveis. O termo de cobertura, pelo argumento de (b), é **quase insensível** a aumentos pequenos de $P^*(x)$ em modas ignoradas. Diante dessa assimetria, o caminho de menor custo durante o treinamento é *melhorar a qualidade nas modas já cobertas* (penalidade alta e gradiente forte) em vez de *cobrir novas modas* (penalidade baixa e gradiente fraco), produzindo *mode collapse*.