Instruções: Responda às questões de múltipla escolha assinalando apenas uma alternativa. Nas questões dissertativas, apresente o raciocínio passo a passo, incluindo fórmulas e cálculos intermediários quando solicitado. Mantenha a notação utilizada em aula: $p_X(x)$ para a densidade modelada, $p_Z(z)$ para a distribuição base, f para a transformação inversível com parâmetros θ, e $J = \partial f/\partial z$ para a Jacobiana. Em camadas *coupling*, $\mathbf{x}^A$ é a metade copiada e $\mathbf{x}^B$ é a metade transformada. Questões marcadas com ☠ são consideradas mais difíceis.

Questões de Múltipla Escolha.

I) Quatro funções $f : \mathbb{R}^d \to \mathbb{R}^d$ são candidatas a compor uma camada de *normalizing flow*. Para que a camada possa ser treinada por máxima verossimilhança e amostrada pela inversa, f deve satisfazer simultaneamente um conjunto específico de propriedades. Qual conjunto abaixo é o mínimo necessário?

(a) Inversível, diferenciável, com Jacobiana cujo determinante seja eficiente de computar, e paramétrica.

(b) Linear por partes, contínua, com determinante igual a 1, e paramétrica.

(c) Estritamente convexa, suave, com Jacobiana ortogonal, e paramétrica.

(d) Inversível, monótona crescente em cada coordenada, com gradiente limitado, e linear.

II) Considere uma camada *affine coupling* aplicada a $\mathbf{x} \in \mathbb{R}^6$ com partição $\mathbf{x}^A \in \mathbb{R}^3$ e $\mathbf{x}^B \in \mathbb{R}^3$. A *coupling network* produziu, a partir de $\mathbf{x}^A$, o vetor de log-escala $\mathbf{s} = (0.2,\ -0.5,\ 0.7)$. Qual é o valor de $\log|\det J|$ para essa camada?

(a) 1.4.

(b) 0.4.

(c) ≈ 1.49.

(d) ≈ 0.13.

III) Você precisa de um modelo gerador que, para qualquer ponto x no espaço de dados, devolva o valor exato de $\log p_X(x)$ (e não apenas um limite inferior ou um sinal de discriminador). Considerando as famílias VAE, GAN e *Normalizing Flow*, qual afirmação descreve corretamente o que cada uma oferece nesse aspecto?

(a) VAE fornece $\log p_X(x)$ exato pelo *decoder*; GAN fornece um limite inferior; NF, apenas amostras.

(b) GAN fornece $\log p_X(x)$ exato via discriminador; VAE fornece apenas amostras; NF, um limite inferior.

(c) NF fornece $\log p_X(x)$ exato via mudança de variável; VAE fornece um limite inferior; GAN, apenas amostras.

(d) As três famílias fornecem $\log p_X(x)$ exato, diferindo apenas na qualidade visual das amostras.

IV) Em uma camada *affine coupling* aplicada a $\mathbf{x} = (x^A, x^B) = (2,\ 3)$, a *coupling network* produz $s(x^A) = -0.5$ e $t(x^A) = 1.0$. Aplicando a regra *forward* $\mathbf{y}^B = \mathbf{x}^B \odot \exp(\mathbf{s}) + \mathbf{t}$, $\mathbf{y}^A = \mathbf{x}^A$, qual é o valor de $\mathbf{y} = (y^A, y^B)$? Use $e^{-0.5} \approx 0.607$.

(a) $(2,\ 2.82)$.

(b) $(2,\ 1.82)$.

(c) $(2,\ 5.95)$.

(d) $(2,\ 3.50)$.

V) Em uma camada *coupling*, a Jacobiana tem a forma por blocos

$$Df = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \dfrac{\partial \hat{f}}{\partial \mathbf{x}^A} & \dfrac{\partial \hat{f}}{\partial \mathbf{x}^B} \end{bmatrix}.$$

Qual conclusão sobre $\det(Df)$ segue dessa estrutura?

(a) $\det(Df) = \det(\partial \hat{f}/\partial \mathbf{x}^A)$, pois o bloco fora da diagonal domina.

(b) $\det(Df) = \det(\partial \hat{f}/\partial \mathbf{x}^B)$, pois o determinante de uma matriz triangular por blocos é o produto dos determinantes dos blocos diagonais e $\det(\mathbf{I}) = 1$.

(c) $\det(Df) = \det(\partial \hat{f}/\partial \mathbf{x}^A) \cdot \det(\partial \hat{f}/\partial \mathbf{x}^B)$.

(d) $\det(Df) = 1$, pois o bloco superior é a identidade.

VI) A perda treinada por máxima verossimilhança num *flow* tem a forma

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \left[\log \left| \det \tfrac{\partial f(z^{(i)}, \theta)}{\partial z^{(i)}} \right| - \log p_Z(z^{(i)}) \right],$$

onde p_Z é a distribuição base (gaussiana padrão) e $z^{(i)} = f^{-1}(x^{(i)}, \theta)$. Sobre os dois termos, qual afirmação é correta?

(a) O termo $\log|\det J|$ depende de θ; o termo $-\log p_Z(z^{(i)})$ é constante em θ e pode ser omitido do gradiente sem alterar o treino.

(b) Ambos os termos dependem de θ: o primeiro pela Jacobiana da rede e o segundo pelo ponto de avaliação $z^{(i)} = f^{-1}(x^{(i)}, \theta)$.

(c) O termo $-\log p_Z(z^{(i)})$ depende de θ pela parametrização da média e variância da gaussiana base; $\log|\det J|$ permanece constante.

(d) Nenhum dos dois termos depende de θ, pois ambos são avaliados em pontos $z^{(i)}$ fixados antes do treinamento começar.

VII) Uma camada *affine coupling* produziu $\mathbf{y} = (y^A, y^B) = (1,\ 2.20)$ a partir da entrada $\mathbf{x}$. A *coupling network* (reavaliada em $\mathbf{x}^A = y^A = 1$) devolve $s = 0.3$ e $t = -0.5$. Use $e^{-0.3} \approx 0.741$. Qual é o valor recuperado de $\mathbf{x}^B$ pela inversa?

(a) 2.00.

(b) 1.26.

(c) 3.64.

(d) 1.63.

VIII) ☠️ Seja $X \sim \mathcal{N}(0, 1)$ e $Y = 2X + 3$. Pela fórmula de mudança de variável, qual é o valor de $p_Y(y)$ em $y = 3$? Use $1/\sqrt{2\pi} \approx 0.399$.

(a) ≈ 0.399.

(b) ≈ 0.200.

(c) ≈ 0.798.

(d) ≈ 0.133.

IX) Em um *Real NVP*, intercala-se **permutações** (ou *linear flows* estruturados) entre camadas *coupling* sucessivas. Qual é o motivo principal?

(a) Reduzir o custo computacional do cálculo do determinante, que cresceria com a profundidade.

(b) Forçar a Jacobiana global a ser ortogonal, o que zera o termo de log-determinante na perda.

(c) Estabilizar o treino normalizando a magnitude dos vetores **s** e **t** produzidos pela *coupling network*.

(d) Garantir que todas as dimensões da entrada sejam efetivamente transformadas ao longo da rede, e não apenas a metade $\mathbf{x}^B$ de uma única partição.

X) Em um *linear flow* $f(\mathbf{x}) = \theta\mathbf{x} + \mathbf{b}$ com $\mathbf{x} \in \mathbb{R}^d$ e θ densa, calcular $\det(\theta)$ e θ^{-1} a cada passo de treino custa $O(d^3)$. Estruturas como decomposição LU ou convoluções 1×1 (Glow) são adotadas porque:

(a) Aumentam a expressividade da camada, permitindo que ela modele relações não lineares entre as dimensões.

(b) Garantem que o determinante seja sempre igual a 1, eliminando o termo $\log|\det J|$ da perda.

(c) Reduzem o custo de determinante e inversa para $O(d)$ ou $O(d^2)$, mantendo a inversibilidade exigida pelo *flow*.

(d) Eliminam a necessidade de uma distribuição base p_Z, pois a transformação fica isométrica.

Gabarito - Objetivas

I) a

II) b

III) c

IV) a

V) b

VI) b

VII) a

VIII) b

IX) d

X) c

Questões Dissertativas.

I) (**Derivação da NLL para *flows***) Partindo do *framework* de máxima verossimilhança

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{N} p_X\left(x^{(i)} \mid \theta\right),$$

derive a perda treinada em *normalizing flows* em três etapas explícitas:

(a) Transforme o arg max do produto em arg min de uma soma negativa de log-verossimilhanças. Justifique em uma frase a vantagem numérica dessa troca.

(b) Substitua $p_X(x^{(i)} \mid \theta)$ usando a fórmula de mudança de variável $p_X(x) = p_Z(z)\,|\partial f/\partial z|^{-1}$, com $z = f^{-1}(x, \theta)$.

(c) Apresente a forma final como soma de dois termos, explicando como cada um deles depende de θ. Comente o papel intuitivo do termo $\log|\det J|$ (qual comportamento da rede ele penaliza).

II) (**Camada *affine coupling* em 4D**) Considere $\mathbf{x} = (x_1, x_2, x_3, x_4) = (0.5,\ 1.0,\ -1.0,\ 2.0)$, com $\mathbf{x}^A = (x_1, x_2)$ e $\mathbf{x}^B = (x_3, x_4)$. A *coupling network* produz, a partir de $\mathbf{x}^A$, os vetores $\mathbf{s} = (0.1,\ -0.2)$ e $\mathbf{t} = (0.3,\ -0.4)$.

(a) Calcule o *forward pass* $\mathbf{y} = f(\mathbf{x})$ usando $\mathbf{y}^B = \mathbf{x}^B \odot \exp(\mathbf{s}) + \mathbf{t}$ e $\mathbf{y}^A = \mathbf{x}^A$. Use $e^{0.1} \approx 1.105$ e $e^{-0.2} \approx 0.819$.

(b) Calcule $\log|\det J|$ para essa camada. Mostre que o resultado depende apenas de $\mathbf{s}$.

(c) Recupere $\mathbf{x}^B$ aplicando a inversa $\mathbf{x}^B = (\mathbf{y}^B - \mathbf{t}) \odot \exp(-\mathbf{s})$. Confirme que os valores originais $(-1.0,\ 2.0)$ são reconstruídos.

(d) ☠ Suponha que a *coupling network* fosse uma rede neural arbitrária, possivelmente não inversível (por exemplo, com camadas ReLU que descartam informação). Explique por que isso **não** compromete a inversibilidade da camada *coupling* como um todo. Em que sentido o ”preço”da inversibilidade global recai apenas sobre a função elemento a elemento $\hat{f}$, e não sobre a sub-rede que produz $(\mathbf{s}, \mathbf{t})$?

III) (**NF, VAE e GAN no trilema**) Considere três eixos para avaliar uma família de modelos geradores: (i) *avaliação* de $p_X(x)$ para um ponto dado, (ii) custo computacional de *amostragem*, (iii) *qualidade visual* típica em datasets de imagens naturais. Para cada um dos três modelos abaixo, indique em que ponto do espaço (eixos i, ii, iii) ele se posiciona e justifique a partir do mecanismo de treino e da arquitetura:

(a) GAN.

(b) VAE.

(c) *Normalizing Flow*.

Conclua identificando qual canto do trilema é ocupado pelos NFs e cite uma aplicação prática em que esse canto é preferível ao dos GANs ou modelos difusores.

IV) ☠ (**Mudança de variável 1D**) Seja $X \sim \mathcal{N}(0, 1)$ e defina $Z = \exp(X)$. O objetivo é derivar a densidade $p_Z(z)$.

(a) Escreva $p_X(x)$ explicitamente.

(b) Determine a função inversa $x = f^{-1}(z)$ e calcule $|dx/dz|$.

(c) Aplique a fórmula $p_Z(z) = p_X(f^{-1}(z))\,|dx/dz|$ e escreva $p_Z(z)$ em forma fechada. Indique o suporte ($z > 0$) e o nome dessa distribuição.

(d) Calcule numericamente $p_Z(z = 1)$. Use $1/\sqrt{2\pi} \approx 0.399$.

Gabarito

Múltipla Escolha.

I) **(a)** As quatro propriedades essenciais discutidas em aula são: paramétrica (para haver o que aprender), inversível (para suportar amostragem e avaliação), diferenciável (para *backprop*) e Jacobiana com determinante eficiente (para tornar a NLL tratável). A alternativa (b) impõe det = 1, perdendo a deformação de volume que é justamente o que carrega a densidade. (c) e (d) acrescentam restrições gratuitas (ortogonalidade, monotonicidade coordenada a coordenada, linearidade) que limitam fortemente a expressividade.

II) **(b)** Para *affine coupling*, $\log|\det J| = \sum_i s_i = 0.2 + (-0.5) + 0.7 = 0.4$. A alternativa (a) seria $\sum_i |s_i| = 1.4$, um erro comum de quem aplica módulo a cada s_i. A (c) corresponde a $\prod_i \exp(s_i) \approx \exp(0.4) \approx 1.49$, ou seja, devolve $|\det J|$ em vez de seu log. A (d) é a média dos s_i.

III) **(c)** *Normalizing flows* fornecem $\log p_X(x)$ exato porque a mudança de variável devolve uma fórmula fechada em função de $p_Z(z)$ e do log-determinante. VAEs maximizam o ELBO, que é apenas um limite inferior de $\log p_X(x)$, não o valor exato. GANs treinam por jogo entre gerador e discriminador e não dão acesso a uma densidade.

IV) **(a)** $y^A = x^A = 2$. $y^B = 3 \cdot e^{-0.5} + 1 \approx 3 \cdot 0.607 + 1 = 1.82 + 1 = 2.82$. A (b) esquece o termo $+t$ ($3 \cdot 0.607 \approx 1.82$). A (c) usa $\exp(+0.5) \approx 1.649$ no lugar de $\exp(-0.5)$, dando $3 \cdot 1.649 + 1 \approx 5.95$. A (d) aplica $x^B + s + t = 3 - 0.5 + 1 = 3.5$, somando em vez de escalar por $\exp(s)$.

V) **(b)** Para uma matriz triangular inferior por blocos $\begin{bmatrix} A & 0 \\ C & D \end{bmatrix}$, o determinante é $\det(A)\det(D)$. Como o bloco superior é $\mathbf{I}$ (com det = 1), sobra $\det(\partial \hat{f}/\partial \mathbf{x}^B)$. O bloco fora da diagonal $\partial \hat{f}/\partial \mathbf{x}^A$ é exatamente o que *não importa* para o determinante. As alternativas (a) e (c) trocam essa propriedade.

VI) **(b)** Ambos os termos recebem gradiente. O $\log|\det J|$ depende de θ diretamente, pela Jacobiana da rede, e penaliza camadas que contraem volume de forma exagerada para "trapacear" empurrando toda a massa para uma região pequena. O termo $-\log p_Z(z)$ tem forma fechada fixa (gaussiana padrão), mas o ponto de avaliação $z^{(i)} = f^{-1}(x^{(i)}, \theta)$ se move quando θ muda; é esse termo que puxa os latentes para a região de alta densidade da base. A (a) é o erro clássico: se o termo da base fosse omitido do gradiente, a rede minimizaria a perda apenas contraindo volume, sem ajustar a distribuição aos dados. A (c) inverte os papéis, a base é fixa e a Jacobiana não é constante. A (d) é falsa, $z^{(i)}$ é recalculado a cada passo de treino.

VII) **(a)** $x^B = (y^B - t) \cdot e^{-s} = (2.20 - (-0.5)) \cdot e^{-0.3} = 2.70 \cdot 0.741 \approx 2.00$. A (b) usa $(y^B - 0.5) \cdot 0.741 = 1.70 \cdot 0.741 \approx 1.26$, esquecendo que $t = -0.5$ e portanto $y^B - t = y^B + 0.5$. A (c) aplica e^{+s} em vez de e^{-s}, dando $2.70 \cdot 1.350 \approx 3.64$. A (d) usa $y^B \cdot e^{-s} = 2.20 \cdot 0.741 \approx 1.63$, omitindo t por completo.

VIII) **(b)** $Y = aX + b$ com $a = 2$, $b = 3$ é a transformação afim padrão. Pela mudança de variável, $p_Y(y) = p_X((y-b)/a) \cdot |1/a|$. Em $y = 3$, $(y-b)/a = 0$, então $p_X(0) = 1/\sqrt{2\pi} \approx 0.399$. Multiplicando por $|1/a| = 1/2$, $p_Y(3) \approx 0.200$. A (a) esquece o fator Jacobiano $1/|a|$ e devolve $p_X(0)$ direto. A (c) multiplica por $|a| = 2$ no lugar de dividir, obtendo ≈ 0.798. A (d) divide por $|a + b| = 3$ no lugar de $|a|$, confundindo a estrutura da fórmula.

IX) **(d)** Uma única camada *coupling* mantém $\mathbf{x}^A$ idêntico, então nada acontece com essas dimensões. Alternar permutações entre camadas faz com que diferentes subconjuntos de coordenadas alternem entre o papel de "A"(copiadas) e "B"(transformadas), garantindo que toda dimensão seja eventualmente afetada. As demais alternativas descrevem efeitos que não correspondem ao papel da permutação.

X) **(c)** A inversibilidade é uma exigência da família *flow*, então estruturas como LU ou convoluções 1×1 não a comprometem. O ganho é puramente computacional: matrizes estruturadas têm determinante e inversa em tempo subcúbico (LU custa $O(d)$ no determinante depois de fatorada; 1×1 convolução tem custo proporcional ao número de canais). A (a) é falsa, transformações lineares continuam lineares mesmo estruturadas. A (b) zera o termo da NLL e perde expressividade. A (d) confunde com mapas isométricos.

Dissertativas.

I) (a) Tomando o logaritmo e trocando o sinal, $\hat{\theta} = \arg\min_\theta \sum_{i=1}^N \left[-\log p_X(x^{(i)} \mid \theta)\right]$. A vantagem é numérica: o produto de N probabilidades pequenas sofre *underflow* rapidamente em ponto flutuante; somar logaritmos é estável e produz gradientes bem condicionados.

(b) Pela mudança de variável, $p_X(x) = p_Z(f^{-1}(x,\theta))\,|\partial f/\partial z|^{-1}$. Aplicando $-\log$, obtém-se $-\log p_X(x) = -\log p_Z(z) + \log|\partial f/\partial z|$, com $z = f^{-1}(x,\theta)$.

(c) Soma final:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \left[\log\left|\frac{\partial f(z^{(i)},\theta)}{\partial z^{(i)}}\right| - \log p_Z(z^{(i)})\right].$$

O termo $\log|\det J|$ depende de θ diretamente (via a Jacobiana da rede f) e penaliza configurações que comprimem volume excessivamente, que de outra forma deixariam a verossimilhança artificialmente alta. O termo $-\log p_Z(z)$ tem densidade em forma fechada e fixa (gaussiana padrão, em geral), mas também depende de θ, através do ponto de avaliação $z = f^{-1}(x,\theta)$; é ele que puxa os latentes para a região de alta densidade da base.

II) (a) $\mathbf{y}^A = (0.5,\ 1.0)$. Para $\mathbf{y}^B$: $y_3 = x_3 \cdot e^{s_1} + t_1 = (-1.0)(1.105) + 0.3 = -1.105 + 0.3 = -0.805$. $y_4 = x_4 \cdot e^{s_2} + t_2 = (2.0)(0.819) + (-0.4) = 1.638 - 0.4 = 1.238$. Logo $\mathbf{y} \approx (0.5,\ 1.0,\ -0.805,\ 1.238)$.

(b) $\log|\det J| = \sum_i s_i = 0.1 + (-0.2) = -0.1$. Como a Jacobiana é triangular por blocos com bloco superior $\mathbf{I}$ e bloco inferior-direito diagonal $\mathrm{diag}(e^{s_1}, e^{s_2})$, o determinante é $\prod_i e^{s_i}$ e seu log é $\sum_i s_i$. Os elementos do bloco inferior-esquerdo (derivadas de $\hat{f}$ em $\mathbf{x}^A$ via $\mathbf{s}$ e $\mathbf{t}$) não aparecem.

(c) $x_3 = (y_3 - t_1)\cdot e^{-s_1} = (-0.805 - 0.3)\cdot e^{-0.1} = (-1.105)(0.905) \approx -1.000$. $x_4 = (y_4 - t_2)\cdot e^{-s_2} = (1.238 - (-0.4))\cdot e^{0.2} = (1.638)(1.221) \approx 2.000$. Os valores originais são reconstruídos, confirmando que a inversa é exata.

(d) A inversibilidade da camada *coupling* depende de duas coisas: $\mathbf{x}^A$ ser copiada para $\mathbf{y}^A$ (portanto recuperável trivialmente) e $\hat{f}(\cdot \mid \theta(\mathbf{x}^A))$ ser inversível como função de $\mathbf{x}^B$ **dado** $\theta(\mathbf{x}^A)$. A sub-rede que produz $(\mathbf{s}, \mathbf{t})$ a partir de $\mathbf{x}^A$ *nunca é invertida*, ela é apenas reavaliada na passagem inversa (note que $\mathbf{x}^A = \mathbf{y}^A$ está disponível em ambas as direções). Assim, o "preço"da inversibilidade global recai apenas sobre a forma fechada de $\hat{f}$ (no caso afim, exponencial mais translação, trivialmente inversível), e a *coupling network* pode ser qualquer MLP/CNN com ReLU, *batch norm*, etc.

III) (a) **GAN**: (i) avaliação inviável (não há densidade explícita, apenas um discriminador que aproxima razões de probabilidades); (ii) amostragem rápida (uma passagem pelo gerador); (iii) qualidade visual historicamente alta em imagens naturais, ao custo de *mode collapse* e cobertura ruim da distribuição.

(b) **VAE**: (i) avaliação aproximada via ELBO (limite inferior de $\log p_X(x)$, não o valor exato); (ii) amostragem rápida (uma passagem pelo *decoder*); (iii) qualidade visual em geral inferior à de GANs e modelos difusores, com tendência a amostras borradas devido à perda de reconstrução pixel a pixel.

(c) **Normalizing Flow**: (i) avaliação exata de $\log p_X(x)$ pela fórmula de mudança de variável; (ii) amostragem rápida (uma passagem direta pelo *flow*); (iii) qualidade visual inferior a GANs e modelos difusores em *datasets* grandes como ImageNet/FFHQ, em parte pela restrição arquitetural (inversibilidade) que limita a expressividade por parâmetro.

Conclusão: NFs ocupam o canto *velocidade + cobertura* (com avaliação exata) do trilema. Aplicações em que esse canto é preferível: detecção de anomalias (precisamos de $\log p_X(x)$ para definir limiar), inferência variacional (substituir $q(z \mid x)$ por uma distribuição expressiva e ainda tratável), física computacional e *Boltzmann generators* (amostragem em equilíbrio com densidade exata para reponderação).

IV) (a) $p_X(x) = \dfrac{1}{\sqrt{2\pi}} \exp\left(-\dfrac{x^2}{2}\right)$.

(b) Como $z = \exp(x)$, então $x = \ln(z)$, definido para $z > 0$. A derivada é $dx/dz = 1/z$, e como $z > 0$, $|dx/dz| = 1/z$.

(c) Aplicando a fórmula:

$$p_Z(z) = p_X(\ln z) \cdot \frac{1}{z} = \frac{1}{z\sqrt{2\pi}} \exp\left(-\frac{(\ln z)^2}{2}\right), \quad z > 0.$$

Essa é a distribuição **log-normal** (com parâmetros $\mu = 0$, $\sigma = 1$).

(d) Em $z = 1$: $\ln(1) = 0$, então $\exp(-0/2) = 1$ e $p_Z(1) = (1/1) \cdot (1/\sqrt{2\pi}) \cdot 1 \approx 0.399$.