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 dos slides: $g[x, \phi]$ para o *encoder* (parâmetros ϕ), $f[z, \theta]$ para o *decoder* (parâmetros θ), $q(z \mid x, \phi)$ para o posterior aproximado e $p(z) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ para o *prior*, conforme apresentadas em aula. Questões marcadas com ☠ são consideradas difíceis.

Questões de Múltipla Escolha.

I) Um pesquisador treina um *autoencoder* sobre imagens 28×28 (entrada com 784 *features*) e observa erro de reconstrução próximo de zero, mas as representações latentes não capturam estrutura semântica útil. Ao inspecionar a arquitetura, ele percebe que o vetor latente possui dimensão 1024 e que todas as camadas usam ativações lineares. Qual é a causa mais provável do problema?

(a) A função de perda escolhida não é compatível com imagens em escala de cinza.

(b) O *decoder* possui parâmetros insuficientes para reconstruir a entrada.

(c) Não há gargalo de informação, então a rede aprende a copiar a entrada.

(d) A inicialização dos pesos foi feita com valores muito pequenos, causando colapso.

II) Em um *denoising autoencoder*, qual é o procedimento correto de treinamento por amostra?

(a) Corromper x gerando $\tilde{x}$, passar $\tilde{x}$ pela rede e comparar a saída com $\tilde{x}$.

(b) Corromper x gerando $\tilde{x}$, passar $\tilde{x}$ pela rede e comparar a saída com x original.

(c) Passar x pela rede e comparar a saída com uma versão corrompida $\tilde{x}$ do mesmo x.

(d) Passar x pela rede e adicionar ruído na representação latente antes do *decoder*.

III) Considere um *autoencoder* vanilla treinado para minimizar o erro de reconstrução. Por que esse modelo, sozinho, não pode ser usado para gerar amostras novas e plausíveis, ainda que reconstrua bem o conjunto de treino?

(a) O *decoder* é determinístico e, portanto, não admite a injeção de ruído necessária para variabilidade.

(b) Não há distribuição conhecida sobre o latente, amostras aleatórias caem fora do suporte aprendido.

(c) O erro de reconstrução tende a zero, o que elimina qualquer diversidade na saída do *decoder*.

(d) O *encoder* não é diferenciável em relação a ϕ, impedindo a propagação do gradiente até a entrada.

IV) O *encoder* de um VAE com latente bidimensional ($D_z = 2$) produz, para uma entrada x, os parâmetros $\mu = (2, 0)$ e $\Sigma = \mathrm{diag}(1, 1)$. Considerando $p(z) = \mathcal{N}(\mathbf{0}, \mathbf{I})$, qual é o valor de $D_{KL}[\, q(z \mid x, \phi) \,\|\, p(z) \,]$?

(a) 0.

(b) 1.

(c) 2.

(d) 4.

V) Durante o *forward pass* de um VAE com latente bidimensional, o *encoder* produz $\mu = (1, -2)$ e desvios-padrão $\sigma = (0.5, 1.5)$ (assumindo Σ diagonal). Foi amostrado $\epsilon^* = (0.4, -0.2)$ a partir de $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Qual é o valor de z^* produzido pelo *reparametrization trick*?

(a) $(1.2, -2.3)$.

(b) $(1.4, -2.2)$.

(c) $(0.9, -1.7)$.

(d) $(1.5, -1.7)$.

VI) Por que o *reparametrization trick* é necessário durante o treinamento de um VAE?

(a) Para garantir que o *prior* $p(z)$ seja exatamente $\mathcal{N}(\mathbf{0}, \mathbf{I})$ durante toda a otimização.

(b) Para que a amostragem de z^* permaneça no caminho diferenciável em relação aos parâmetros do *encoder*.

(c) Para reduzir o custo computacional da estimativa *Monte Carlo* do termo de reconstrução.

(d) Para forçar a distribuição posterior $q(z \mid x, \phi)$ a ter média zero e variância unitária.

VII) O ELBO usado no treinamento de um VAE pode ser escrito como a soma de dois termos. Qual alternativa identifica corretamente o papel de cada termo na função objetivo?

(a) Termo de reconstrução: $-D_{KL}[q(z \mid x, \phi) \,\|\, p(z)]$. Termo de regularização: $\mathbb{E}_q[\log p(x \mid z, \theta)]$.

(b) Termo de reconstrução: $\mathbb{E}_q[\log p(x \mid z, \theta)]$. Termo de regularização: $-D_{KL}[q(z \mid x, \phi) \,\|\, p(z)]$.

(c) Termo de reconstrução: $\mathbb{E}_{p(z)}[\log p(x \mid z, \theta)]$. Termo de regularização: $D_{KL}[p(z) \,\|\, q(z \mid x, \phi)]$.

(d) Termo de reconstrução: $\log p(x \mid \theta)$. Termo de regularização: $\mathbb{E}_q[\log q(z \mid x, \phi)]$.

VIII) Após treinar um VAE com *decoder* muito expressivo, observa-se que $D_{KL}[q(z \mid x, \phi) \,\|\, p(z)]$ tende a zero para todas as entradas, enquanto o erro de reconstrução do *decoder* permanece elevado. O que esse padrão indica?

(a) A taxa de aprendizado está alta demais e o gradiente do termo KL está dominando o gradiente da reconstrução.

(b) O *encoder* aprendeu a igualar $q(z \mid x, \phi)$ ao *prior* e o latente carrega pouca informação sobre x.

(c) O ruído ϵ^* do *reparametrization trick* não está sendo amostrado de forma independente para cada exemplo.

(d) A dimensão do latente D_z está pequena demais para representar a distribuição dos dados.

IX) Depois de treinar um VAE, deseja-se gerar uma nova amostra x^* que se assemelhe aos dados de treino. Qual procedimento corresponde à amostragem ancestral correta?

(a) Tomar um x qualquer do conjunto de treino, passar pelo *encoder* para obter μ, Σ e produzir x^* via *decoder* aplicado em μ.

(b) Amostrar $z^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ e passar z^* pelo *decoder* para obter os parâmetros de $p(x \mid z^*, \theta)$.

(c) Amostrar z^* uniformemente em um cubo de aresta 1 centrado na origem e passar pelo *decoder*.

(d) Amostrar z^* de $q(z \mid x, \phi)$ com x tirado do conjunto de treino e copiar a saída do *decoder*.

X) ☠ Considere a fronteira entre um *autoencoder* clássico e um *variational autoencoder*. Qual afirmação abaixo descreve mais precisamente o papel da inferência variacional no VAE, dado que $p(x \mid \theta) = \int p(x \mid z, \theta)\, p(z)\, dz$ é, em geral, intratável?

(a) A inferência variacional substitui o termo $\log p(x \mid \theta)$ por um limite inferior tratável envolvendo $q(z \mid x, \phi)$, que é otimizado em conjunto com θ.

(b) A inferência variacional fornece uma forma exata de calcular $p(x \mid \theta)$ via uma única amostra de z, dispensando integração numérica.

(c) A inferência variacional transforma o problema em uma classificação binária entre amostras reais e amostras geradas pelo *decoder*.

(d) A inferência variacional garante que o *encoder* aprenda exatamente a distribuição posterior verdadeira $p(z \mid x, \theta)$.

XI) ☠ Considere dois *encoders* A e B de um VAE com latente bidimensional, ambos produzindo $\mu = (0, 0)$. O *encoder* A devolve $\Sigma_A = \mathrm{diag}(0.01, 0.01)$, enquanto o B devolve $\Sigma_B = \mathrm{diag}(1, 1)$. Comparando $D_{KL}[q_A \,\|\, p(z)]$ e $D_{KL}[q_B \,\|\, p(z)]$ contra $p(z) = \mathcal{N}(\mathbf{0}, \mathbf{I})$, qual afirmação está correta?

(a) $D_{KL}^A < D_{KL}^B$, pois variâncias menores aproximam q de uma massa pontual e reduzem a divergência.

(b) $D_{KL}^A = D_{KL}^B = 0$, pois ambas as médias coincidem com a média do *prior*.

(c) $D_{KL}^A > D_{KL}^B$, pois o termo $-\log\det\Sigma$ cresce quando as variâncias se aproximam de zero.

(d) $D_{KL}^A \approx D_{KL}^B$, pois o termo dominante é $\mu^\top \mu$, que se anula em ambos os casos.

XII) Um *autoencoder esparso* é uma variante em que a dimensão do latente pode ser *maior* que a da entrada (overcomplete), mas ainda assim a rede aprende representações úteis. Como essa configuração evita a solução trivial de cópia?

(a) Penalizando a ativação do latente para que poucas unidades fiquem ativas simultaneamente.

(b) Aplicando *dropout* apenas no *decoder* para impedir a passagem direta da informação.

(c) Reduzindo a taxa de aprendizado sempre que o erro de reconstrução fica abaixo de um limiar.

(d) Substituindo o erro quadrático médio por entropia cruzada categórica no *decoder*.

Gabarito - Objetivas

I) c

II) b

III) b

IV) c

V) a

VI) b

VII) b

VIII) b

IX) b

X) a

XI) c

XII) a

Resoluções - Objetivas

I) **(c)** Com $D_z = 1024 > 784$ e ativações lineares em todas as camadas, não existe gargalo de informação: a rede pode implementar a identidade (ou um par de mapas lineares que se cancelam) e copiar a entrada, zerando o erro sem aprender estrutura. A (a) é irrelevante para o sintoma descrito. A (b) contradiz o erro de reconstrução próximo de zero. A (d) produziria erro de reconstrução alto, o oposto do observado.

II) **(b)** O *denoising autoencoder* corrompe a entrada (gerando $\tilde{x}$) mas usa o x limpo como alvo, forçando a rede a aprender a remover o ruído. A (a) treina a rede a reproduzir a própria entrada corrompida, sem efeito de limpeza. A (c) inverte os papéis e ensinaria a rede a adicionar ruído. A (d) descreve perturbação no latente, que não caracteriza o *denoising autoencoder*.

III) **(b)** Sem uma distribuição conhecida sobre o latente, não há como escolher z para alimentar o *decoder*: pontos sorteados ao acaso tendem a cair fora da região efetivamente ocupada pelas codificações do treino, e o *decoder* devolve saídas implausíveis. A (a) não é o impedimento, o *decoder* de um VAE também é determinístico dado z. A (c) é um *non sequitur*, reconstruir bem não elimina diversidade. A (d) parte de premissa falsa, o *encoder* é diferenciável.

IV) **(c)** $D_{KL} = \frac{1}{2}\left(\mathrm{Tr}[\Sigma] + \mu^\top\mu - D_z - \log\det\Sigma\right) = \frac{1}{2}(2 + 4 - 2 - 0) = 2$. A (a) corresponde a ignorar a média (com $\mu = \mathbf{0}$ e $\Sigma = \mathbf{I}$ a divergência seria nula). A (b) corresponde a usar $\|\mu\| = 2$ sem elevar ao quadrado. A (d) corresponde a usar $\mu^\top\mu = 4$ sozinho, sem o fator $\frac{1}{2}$ e sem os demais termos.

V) **(a)** $z^* = \mu + \sigma \odot \epsilon^* = (1 + 0.5 \cdot 0.4,\ -2 + 1.5 \cdot (-0.2)) = (1.2,\ -2.3)$. A (b) soma ϵ^* diretamente, sem escalar por σ. As alternativas (c) e (d) resultam de trocas de componentes ou de sinais na aplicação elemento a elemento.

VI) **(b)** O nó de amostragem é estocástico e não admite derivada em relação a ϕ. Reescrevendo $z^* = \mu + \Sigma^{1/2}\epsilon^*$, a aleatoriedade fica isolada em ϵ^* (amostrado de uma distribuição fixa) e o caminho por μ e $\Sigma^{1/2}$ permanece diferenciável. A (a) confunde o *prior*, fixado por hipótese, com algo imposto pelo truque. A (c) não procede, o custo da estimativa não muda. A (d) descreveria um colapso forçado do posterior, que não é o objetivo.

VII) **(b)** O ELBO é $\mathbb{E}_q[\log p(x \mid z, \theta)] - D_{KL}[q(z \mid x, \phi) \,\|\, p(z)]$: o primeiro termo mede a qualidade da reconstrução, o segundo regulariza q na direção do *prior*. A (a) troca os papéis dos dois termos. A (c) usa a esperança sob o *prior* e inverte os argumentos da KL. A (d) confunde o ELBO com a própria verossimilhança marginal, que é intratável.

VIII) **(b)** KL próxima de zero com reconstrução ruim é sinal do *posterior collapse*: $q(z \mid x, \phi) \approx p(z)$ para todo x e o latente deixa de informar o *decoder*. A (a) descreveria instabilidade geral de otimização, não esse padrão específico. A (c) não produz esse sintoma. A (d) implicaria KL tipicamente alta, não nula.

IX) **(b)** A amostragem ancestral segue o modelo gerador: $z^* \sim p(z) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ e, em seguida, o *decoder* produz os parâmetros de $p(x \mid z^*, \theta)$. A (a) descreve reconstrução de um dado existente, não geração. A (c) amostra da distribuição errada. A (d) condiciona em um exemplo do treino, tendendo a reproduzi-lo em vez de gerar algo novo.

X) **(a)** Como $p(x \mid \theta)$ envolve uma integral intratável, a inferência variacional introduz $q(z \mid x, \phi)$ e otimiza o ELBO, um limite inferior tratável de $\log p(x \mid \theta)$, conjuntamente em ϕ e θ. A (b) superestima o que uma única amostra de z fornece. A (c) descreve uma GAN. A (d) é falsa, q apenas aproxima o posterior verdadeiro.

XI) **(c)** $D_{KL}^{A} = \frac{1}{2}\left(0.02 + 0 - 2 - \ln(10^{-4})\right) = \frac{1}{2}(0.02 - 2 + 9.21) \approx 3.6$, enquanto $D_{KL}^{B} = \frac{1}{2}(2 + 0 - 2 - 0) = 0$. O termo $-\log\det\Sigma$ explode quando as variâncias tendem a zero, logo $D_{KL}^{A} > D_{KL}^{B}$. A (a) inverte a intuição, uma massa quase pontual está longe de uma gaussiana de variância unitária. A (b) considera apenas as médias e ignora as covariâncias. A (d) ignora o termo de determinante, que é justamente o dominante aqui.

XII) **(a)** A penalidade de esparsidade força poucas unidades ativas por instância, criando um gargalo de capacidade efetiva mesmo com latente *overcomplete*: a rede precisa selecionar o que representar. A (b) não impede a cópia pelo restante das conexões. A (c) e a (d) alteram a dinâmica de otimização e a função de perda, mas nenhuma das duas remove a solução de identidade.

Questões Dissertativas.

I) Considere um *autoencoder* aplicado a imagens em escala de cinza com $D_x = 784$ *pixels* (entrada vetorizada de 28×28). Para cada uma das configurações abaixo, indique se o modelo é *undercomplete* ou *overcomplete*, e discuta brevemente o efeito esperado sobre a representação aprendida.

(a) Latente com $D_z = 32$ unidades, sem outras regularizações.

(b) Latente com $D_z = 2000$ unidades, sem outras regularizações.

(c) Latente com $D_z = 2000$ unidades, treinado com penalidade de esparsidade na ativação do latente.

II) Seja um VAE com latente $D_z = 2$. Para uma entrada x, o *encoder* produz $\mu = (1, -1)$ e $\Sigma = \text{diag}(0.5, 2)$. Calcule, passo a passo, $D_{KL}[\, q(z \mid x, \phi) \,\|\, \mathcal{N}(\mathbf{0}, \mathbf{I})\,]$ usando a fórmula fechada apresentada em aula. Em seguida, comente quais dos quatro termos ($\text{Tr}[\Sigma]$, $\mu^\top \mu$, D_z, $\log\det\Sigma$) mais contribuíram para o valor obtido.

III) Explique o *reparametrization trick* respondendo aos itens abaixo.

(a) Qual é o problema do ponto de vista da *backpropagation* ao amostrar $z^* \sim q(z \mid x, \phi)$ diretamente?

(b) Escreva a forma reparametrizada de z^* usada em VAEs com posterior gaussiano de média μ e covariância diagonal Σ, identificando claramente quais variáveis ficam no caminho diferenciável e qual variável é a fonte de estocasticidade.

(c) ☠️ Mostre, usando a forma reparametrizada, que $\mathbb{E}[z^*] = \mu$ e que $\text{Cov}[z^*] = \Sigma$.

IV) Compare um *autoencoder* clássico e um *variational autoencoder* respondendo:

(a) O que cada *encoder* produz para uma mesma entrada x?

(b) Por que somente o VAE permite *amostragem ancestral* de novos x^* após o treinamento?

(c) Qual é a função objetivo otimizada em cada caso e por que o VAE não otimiza diretamente $\log p(x \mid \theta)$?

V) Considere um VAE treinado em MNIST cuja loss de treino apresenta o seguinte comportamento ao longo das épocas: o termo $D_{KL}[q(z \mid x, \phi) \,\|\, p(z)]$ cai rapidamente para perto de zero, enquanto o termo de reconstrução estagna em um valor alto. As amostras geradas a partir de $z^* \sim p(z)$ resultam em imagens de aparência média, com pouca variação entre amostras distintas.

(a) Nomeie o fenômeno observado e explique o que ele significa em termos da informação que o latente z carrega sobre x.

(b) Por que esse fenômeno é especialmente comum quando o *decoder* é muito expressivo (por exemplo, autoregressivo)?

(c) Cite uma estratégia conhecida para mitigar o problema e explique, em uma frase, como ela age sobre o ELBO.

Gabarito

I) Configurações.

(a) *Undercomplete* ($D_z = 32 < 784$). O gargalo força o *encoder* a comprimir, descartando informações redundantes; tende a produzir representações úteis para tarefas posteriores.

(b) *Overcomplete* ($D_z = 2000 > 784$), sem regularização. A rede pode aprender uma cópia trivial ou indexar as instâncias do treino (memorização), com representação latente pouco informativa.

(c) *Overcomplete* ($D_z = 2000$), porém regularizado por esparsidade. A penalidade obriga que apenas poucas unidades do latente fiquem ativas por instância, recuperando a utilidade da representação mesmo sem gargalo dimensional.

II) Cálculo passo a passo.

$$\begin{aligned}
\mathrm{Tr}[\Sigma] &= 0.5 + 2 = 2.5 \\
\mu^\top \mu &= 1^2 + (-1)^2 = 2 \\
\log\det\Sigma &= \log(0.5) + \log(2) \approx -0.693 + 0.693 = 0 \\
D_z &= 2 \\
D_{KL} &= \tfrac{1}{2}\big(\mathrm{Tr}[\Sigma] + \mu^\top\mu - D_z - \log\det\Sigma\big) \\
&= \tfrac{1}{2}\big(2.5 + 2 - 2 - 0\big) = \tfrac{1}{2}\,(2.5) = 1.25.
\end{aligned}$$

Os dois termos que mais contribuem são $\mathrm{Tr}[\Sigma] = 2.5$ (dispersão acima do *prior* unitário, sobretudo pela segunda coordenada com $\sigma^2 = 2$) e $\mu^\top\mu = 2$ (desvio da origem). O termo $\log\det\Sigma$ acaba se anulando porque as variâncias 0.5 e 2 se compensam em escala logarítmica.

III) *Reparametrization trick*.

(a) A operação de amostragem é estocástica e não tem derivada bem definida com respeito aos parâmetros que produzem a distribuição. Sem reparametrização, o gradiente de $\mathcal{L}$ não consegue fluir do *decoder* de volta até ϕ através do nó de amostragem.

(b) Escrevemos $z^* = \mu + \Sigma^{1/2}\,\epsilon^*$, com $\epsilon^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. As quantidades μ e $\Sigma^{1/2}$ (saídas do *encoder*, função de ϕ) permanecem no caminho diferenciável; a estocasticidade fica concentrada em ϵ^*, que é amostrado de uma distribuição fixa e *não* depende de ϕ.

(c) Como $\epsilon^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ é independente de μ e Σ:

$$\begin{aligned}
\mathbb{E}[z^*] &= \mu + \Sigma^{1/2}\,\mathbb{E}[\epsilon^*] = \mu + \mathbf{0} = \mu. \\
\mathrm{Cov}[z^*] &= \mathbb{E}\big[(z^* - \mu)(z^* - \mu)^\top\big] \\
&= \mathbb{E}\big[\Sigma^{1/2}\,\epsilon^*\epsilon^{*\top}\,\Sigma^{1/2}\big] \\
&= \Sigma^{1/2}\,\mathbb{E}[\epsilon^*\epsilon^{*\top}]\,\Sigma^{1/2} \\
&= \Sigma^{1/2}\,\mathbf{I}\,\Sigma^{1/2} = \Sigma.
\end{aligned}$$

Logo z^* é distribuído como $\mathcal{N}(\mu, \Sigma)$, conforme desejado.

IV) Comparação AE versus VAE.

(a) O *encoder* de um AE clássico produz um *ponto* no espaço latente, $z = g[x, \phi]$, determinístico dada a entrada. O *encoder* de um VAE produz os parâmetros de uma *distribuição* sobre o latente, $q(z \mid x, \phi) = \mathcal{N}(\mu, \Sigma)$.

(b) No AE clássico não há uma distribuição conhecida sobre o latente, então amostrar z ao acaso pode cair em regiões sem suporte aprendido pelo *decoder*. No VAE o termo KL aproxima $q(z \mid x, \phi)$ do *prior* $p(z)$, garantindo que amostras $z^* \sim p(z)$ permaneçam em regiões plausíveis para o *decoder*.

(c) O AE clássico minimiza apenas o erro de reconstrução, $J(x, f[g[x, \phi], \theta])$. O VAE maximiza o ELBO, que é um limite inferior de $\log p(x \mid \theta)$. Não otimizamos $\log p(x \mid \theta)$ diretamente porque a marginal $p(x \mid \theta) = \int p(x \mid z, \theta)\, p(z)\, dz$ envolve uma integral intratável devido à não linearidade do *decoder*.

V) Diagnóstico do treino.

(a) O fenômeno é o *posterior collapse*: o *encoder* aprendeu a igualar $q(z \mid x, \phi)$ ao *prior* para qualquer x, de modo que o latente z deixa de carregar informação sobre a entrada. A consequência é que amostras geradas a partir do *prior* se concentram em torno de uma reconstrução média, com pouca diversidade.

(b) Quando o *decoder* é poderoso o bastante para modelar $p(x \mid z, \theta)$ ignorando z (por exemplo, modelos autoregressivos sobre os *pixels*), o otimizador encontra um ótimo barato em zerar o termo KL: a contribuição de z para o ELBO fica marginal e o gradiente empurra q na direção de $p(z)$.

(c) Estratégias comuns incluem *KL annealing* (introduzir o termo KL gradualmente, começando com peso baixo) e *free bits* (não penalizar a KL abaixo de um limiar por dimensão do latente). Ambas reduzem a pressão do termo de regularização nas fases iniciais, permitindo que o *decoder* aprenda a usar z antes que o KL force o colapso.