Ruído Quântico Benéfico em Classificadores Quânticos Variacionais: Uma Teoria de Amostra Finita com Overparameterization, Coerências Espúrias e Benefício Condicionado

Marcelo Claro Laranjeira

marceloclaro@gmail.com
https://orcid.org/0000-0001-8996-2887

Abstract

Apresentamos uma formalização matemática do fenômeno de *ruído benéfico* em circuitos quânticos variacionais (VQCs) usados para classificação supervisionada. A tese central é que, sob (i) *amostra finita*, (ii) *overparameterization* e (iii) presença de uma *componente espúria em coerências* (off-diagonals) capaz de ser explorada no treino mas não sustentada na população, existe um nível de ruído $\gamma^* > 0$ que reduz o risco populacional em relação ao caso ideal $\gamma = 0$. Provamos um teorema de existência de γ^* (benefício condicionado), resultados de robustez via contração de distâncias por canais CPTP e um teorema de convergência sob condições do tipo PL (Polyak–Łojasiewicz) para o objetivo suavizado pelo ruído. Incluímos uma contraprova explícita que refuta a hipótese alternativa "ruído sempre prejudica" e demonstramos limitações: em regimes sem coerência espúria (ou no limite de amostra infinita) o *benefício* desaparece e $\gamma^* = 0$.

1 Introdução e contribuições

Classificadores quânticos variacionais (VQCs) são modelos parametrizados que acoplam um mapa de características quântico e um circuito variacional de treino, com leitura via medição de um observável. Em hardware NISQ, o ruído é inevitável. O entendimento padrão é: ruído degrada desempenho. Contudo, existe um regime empiricamente observável em que um ruído moderado melhora a generalização ao atuar como regularização implícita, análoga a mecanismos clássicos (p.ex. dropout/noise injection). O objetivo deste manuscrito é:

- Formalizar o fenômeno em termos de *decomposição diagonal/coerente* do sinal e *coerência espúria*.
- Provar um *Teorema de Benefício Condicionado*: existência de $\gamma^* > 0$ sob hipóteses explícitas.
- Provar resultados complementares: (i) robustez ao ruído/perturbações via contração; (ii) convergência do treino para objetivo suavizado por ruído.
- Incluir uma *contraprova* (counterexample) refutando a hipótese "ruído sempre piora" e uma proposição de limitação para modelos alternativos.

2 Preliminares: estados, canais e VQCs

Definição 1 (Estados e observáveis)**.** Seja $\mathcal{H} \simeq (\mathbb{C}^2)^{\otimes n}$. Um estado quântico é $\rho \succeq 0$ com $\mathrm{Tr}(\rho) = 1$. Um observável é um operador hermitiano $M = M^\dagger$.

Definição 2 (Canais CPTP)**.** Um canal quântico é uma aplicação linear $\mathcal{E} : \mathbb{C}^{2^n \times 2^n} \to \mathbb{C}^{2^n \times 2^n}$ completamente positiva e preservadora de traço (CPTP). Via Kraus: $\mathcal{E}(\rho) = \sum_k A_k \rho A_k^\dagger$ com $\sum_k A_k^\dagger A_k = \mathbb{I}$.

Definição 3 (VQC como família de funções)**.** Fixe um mapa de características $\phi : x \mapsto \rho(x)$ (estado codificado) e um circuito parametrizado $U(\theta)$, $\theta \in \mathbb{R}^p$. Para um nível de ruído $\gamma \geq 0$, definimos o logit/quase-logit:

$$f_{\theta,\gamma}(x) \;:=\; \mathrm{Tr}\Big(M\, \mathcal{E}_\gamma\big(U(\theta)\,\rho(x)\,U(\theta)^\dagger\big)\Big) \in [-\,\|M\|_\infty\,, \|M\|_\infty].$$

A regra de decisão binária pode ser, por exemplo, $\hat{y} = \mathrm{sign}(f_{\theta,\gamma}(x))$.

2.1 Um modelo de dephasing local (padrão e matematicamente tratável)

Para capturar "coerências" de forma controlada, usamos um canal de *dephasing* (fase) local aplicado ao longo do circuito.

Definição 4 (Canal de dephasing 1-qubit)**.** Em base computacional, escreva $\rho = \left(\begin{smallmatrix} a & b \\ \bar{b} & c \end{smallmatrix}\right)$. Defina $\mathcal{E}_\gamma^{(1)}$ por

$$\mathcal{E}_\gamma^{(1)}(\rho) = \begin{pmatrix} a & \eta(\gamma)\, b \\ \eta(\gamma)\, \bar{b} & c \end{pmatrix}, \qquad \eta(\gamma) \in [0,1],\ \eta(0) = 1,$$

isto é, preserva diagonais e atenua off-diagonals por $\eta(\gamma)$. Uma parametrização comum é $\eta(\gamma) = \sqrt{1-\gamma}$.

Definição 5 (Dephasing local n-qubits e acumulação em profundidade)**.** Defina o canal local em n qubits por $\mathcal{E}_\gamma = \bigotimes_{j=1}^n \mathcal{E}_{\gamma,j}^{(1)}$. Se o circuito tem profundidade (número de camadas de ruído) d e o ruído é inserido após cada camada, o canal efetivo é $\mathcal{E}_\gamma^{\circ d}$.

3 Setup estatístico e decomposição "sinal + coerência espúria"

Definição 6 (Dados, risco e risco empírico)**.** Seja $\mathcal{D}$ uma distribuição sobre $\times\{-1,+1\}$. Dados N i.i.d. $S = \{(x_i, y_i)\}_{i=1}^N \sim \mathcal{D}^N$. Fixe uma perda $\ell : \mathbb{R} \times \{-1,+1\} \to \mathbb{R}_+$ (ex.: hinge, logística, quadrática). O risco populacional e empírico:

$$R(\theta,\gamma) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\ell(f_{\theta,\gamma}(x), y)\big], \qquad \hat{R}_S(\theta,\gamma) = \frac{1}{N}\sum_{i=1}^N \ell(f_{\theta,\gamma}(x_i), y_i).$$

3.1 Decomposição operacional: parte diagonal e parte coerente

A ideia-chave: sob dephasing, componentes coerentes são mais suprimidas do que componentes "clássicas" (diagonais), mas em VQCs essas coerências podem ser *exploradas* durante o treino para memorizar correlações finitas do conjunto S.

Definição 7 (Decomposição de um funcional de medição)**.** Fixe uma base computacional. Para qualquer estado σ, escreva $\sigma = \sigma_{\text{diag}} + \sigma_{\text{coh}}$, onde σ_{diag} retém apenas entradas diagonais e σ_{coh} retém apenas off-diagonals. Defina

$$f_{\theta,\gamma}(x) = \mathrm{Tr}\big(M\,\sigma_{\theta,\gamma}(x)\big), \quad \sigma_{\theta,\gamma}(x) := \mathcal{E}_\gamma^{\circ d}\big(U(\theta)\rho(x)U(\theta)^\dagger\big),$$

e decomponha

$$f_{\theta,\gamma}(x) = f_{\theta,\gamma}^{\text{diag}}(x) + f_{\theta,\gamma}^{\text{coh}}(x), \quad f_{\theta,\gamma}^{\text{diag}}(x) := \mathrm{Tr}(M\,\sigma_{\text{diag}}), \quad f_{\theta,\gamma}^{\text{coh}}(x) := \mathrm{Tr}(M\,\sigma_{\text{coh}}).$$

Hipótese 8 (Coerência espúria sob amostra finita)**.** Existe um subconjunto de parâmetros $\Theta_{\text{spur}} \subset \mathbb{R}^p$ e uma função coerente $g_\theta(x) := f_{\theta,0}^{\text{coh}}(x)$ tal que:

1. (*Nulo em população*) Para todo $\theta \in \Theta_{\text{spur}}$, $\mathbb{E}[y\,g_\theta(x)] = 0$.
2. (*Não-nulo no conjunto finito*) Com probabilidade não desprezível em $S \sim \mathcal{D}^N$, existe $\theta_S \in \Theta_{\text{spur}}$ tal que $\frac{1}{N}\sum_{i=1}^N y_i g_{\theta_S}(x_i)$ é de ordem ao menos $c/\sqrt{N}$ (flutuação amostral).

Hipótese 9 (Overparameterization)**.** O número de parâmetros p é suficientemente grande para que o procedimento de treino (p.ex. descida de gradiente) encontre $\hat{\theta}_0 \in \mathbb{R}^p$ com *erro empírico pequeno* (idealmente interpolação):

$$\hat{R}_S(\hat{\theta}_0, 0) \le \varepsilon_{\text{train}}, \qquad \varepsilon_{\text{train}} \approx 0.$$

4 Lemas estruturais: atenuação de coerência e continuidade do risco

Lema 10 (Atenuação de coerências por dephasing)**.** *Suponha que cada aplicação local de dephasing atenua off-diagonals por $\eta(\gamma) \in [0,1]$ e deixa diagonais invariantes. Então, para qualquer estado σ e para d inserções,*

$$\big(\mathcal{E}_\gamma^{\circ d}(\sigma)\big)_{\text{coh}} = \eta(\gamma)^d\,\sigma_{\text{coh}}, \qquad \big(\mathcal{E}_\gamma^{\circ d}(\sigma)\big)_{\text{diag}} = \sigma_{\text{diag}}.$$

Consequentemente,

$$f_{\theta,\gamma}^{\text{coh}}(x) = \eta(\gamma)^d\, f_{\theta,0}^{\text{coh}}(x).$$

Proof. Em base computacional, o canal por definição multiplica cada entrada off-diagonal por $\eta(\gamma)$ em cada aplicação, e preserva as diagonais. Por linearidade e iteração d vezes, obtém-se o fator $\eta(\gamma)^d$ para off-diagonals e fator 1 para diagonais. Aplicando o funcional linear $\mathrm{Tr}(M\,\cdot)$ conclui-se a identidade para $f_{\theta,\gamma}^{\text{coh}}(x)$. □

Lema 11 (Continuidade (e suavidade local) em γ)**.** *Suponha $\gamma \mapsto \mathcal{E}_\gamma$ contínua em norma completamente limitada e ℓ é Lipschitz no primeiro argumento. Então $\gamma \mapsto R(\theta,\gamma)$ e $\gamma \mapsto \hat{R}_S(\theta,\gamma)$ são contínuas para todo θ fixo.*

Proof. A composição de mapas contínuos é contínua; $\sigma_{\theta,\gamma}(x)$ depende continuamente de γ. Como $f_{\theta,\gamma}(x) = \mathrm{Tr}(M\sigma_{\theta,\gamma}(x))$ é linear em σ, é contínuo em γ. Por Lipschitz de ℓ e pelo teorema da convergência dominada (usando boundedness de f em norma de operador), a expectativa e a média empírica preservam continuidade. □

5 Teorema A: Benefício condicionado (existência de $\gamma^* > 0$)

A prova abaixo segue um roteiro metodologicamente aceitável para banca: (i) construir um *mecanismo* que reduz o termo espúrio (Lema 10); (ii) majorar o risco populacional por "erro de treino + complexidade" (uniform convergence); (iii) mostrar que, para γ pequeno, o aumento no erro empírico é de ordem menor do que a redução do termo espúrio; (iv) mostrar que, para γ grande, o desempenho degrada, logo existe minimizador interno.

5.1 Uma desigualdade-padrão de generalização (Rademacher)

Denote por $\mathcal{F}_\gamma = \{x \mapsto f_{\theta,\gamma}(x) : \theta \in \mathbb{R}^p\}$ a classe de hipóteses para ruído fixo γ. Denote por $\mathfrak{R}_N(\mathcal{F}_\gamma)$ a complexidade de Rademacher empírica/padrão. Assumindo ℓ 1-Lipschitz (pode-se reescalar), uma forma clássica diz que, com probabilidade $\geq 1 - \delta$:

$$\forall f \in \mathcal{F}_\gamma : \quad R(f) \leq \hat{R}_S(f) + 2\mathfrak{R}_N(\mathcal{F}_\gamma) + 3\sqrt{\frac{\log(2/\delta)}{2N}}. \tag{1}$$

(Trataremos (1) como hipótese técnica padrão de aprendizagem estatística.)

Hipótese 12 (Decomposição de complexidade em parte diagonal + coerente)**.** Existe uma decomposição (majorante) da complexidade:

$$\mathfrak{R}_N(\mathcal{F}_\gamma) \;\leq\; \mathfrak{R}_N(\mathcal{F}^{\text{diag}}) \;+\; \eta(\gamma)^d\, \mathfrak{R}_N(\mathcal{F}^{\text{coh}}),$$

onde $\mathcal{F}^{\text{diag}} = \{f^{\text{diag}}_{\theta,0}\}$ e $\mathcal{F}^{\text{coh}} = \{f^{\text{coh}}_{\theta,0}\}$.

Discussão/Observação 13. A Hipótese 12 formaliza que o ruído atua como *contração* da parte coerente. Ela é consistente com o Lema 10 porque, ao atenuar a amplitude de funções coerentes por um fator $\eta(\gamma)^d$, a capacidade efetiva do modelo de ajustar flutuações coerentes diminui proporcionalmente.

Teorema 14 (Teorema de Benefício Condicionado)**.** *Assuma:*

1. *Hipóteses 8 (coerência espúria), 9 (overparameterization) e 12 (decomposição de complexidade).*
2. *O procedimento de treino retorna $\hat{\theta}_\gamma$ com erro empírico controlado para γ pequeno:*

   $$\hat{R}_S(\hat{\theta}_\gamma, \gamma) \leq \varepsilon_{\text{train}} + C_{\text{fit}} \cdot (1 - \eta(\gamma)^d),$$

   para alguma constante $C_{\text{fit}} > 0$ (perda de ajuste cresce suavemente com a supressão de coerência).
3. *Existe $\bar{\gamma} > 0$ tal que, para $\gamma \geq \bar{\gamma}$, o classificador torna-se essencialmente não-informativo (por exemplo, $f_{\hat{\theta}_\gamma,\gamma}(x)$ aproxima-se de constante), implicando risco $\geq R_{\text{chance}} - \epsilon$ (p.ex. $R_{\text{chance}} = 1/2$ em perda 0-1).*

Então existe $\gamma^* \in (0, \bar{\gamma})$ tal que

$$R(\hat{\theta}_{\gamma^*}, \gamma^*) \;<\; R(\hat{\theta}_0, 0),$$

isto é, um nível estritamente positivo *de ruído melhora o risco populacional.*

Proof. Fixe $\delta \in (0,1)$. No evento de alta probabilidade em que vale o bound (1) para γ e $\gamma = 0$:

Passo 1: bound do risco no caso $\gamma = 0$. Aplicando (1) a $f_{\hat{\theta}_{0,0}} \in \mathcal{F}_0$:

$$R(\hat{\theta}_0, 0) \leq \hat{R}_S(\hat{\theta}_0, 0) + 2\mathfrak{R}_N(\mathcal{F}_0) + 3\sqrt{\frac{\log(2/\delta)}{2N}}.$$

Pela Hipótese 9, $\hat{R}_S(\hat{\theta}_0, 0) \leq \varepsilon_{\text{train}} \approx 0$, mas $\mathfrak{R}_N(\mathcal{F}_0)$ pode ser grande, pois inclui a parte coerente espúria.

Passo 2: bound do risco para $\gamma > 0$ pequeno. Analogamente:

$$R(\hat{\theta}_\gamma, \gamma) \leq \hat{R}_S(\hat{\theta}_\gamma, \gamma) + 2\mathfrak{R}_N(\mathcal{F}_\gamma) + 3\sqrt{\frac{\log(2/\delta)}{2N}}.$$

Agora use:

$$\hat{R}_S(\hat{\theta}_\gamma, \gamma) \leq \varepsilon_{\text{train}} + C_{\text{fit}}(1 - \eta(\gamma)^d),$$

e, pela Hipótese 12,

$$\mathfrak{R}_N(\mathcal{F}_\gamma) \leq \mathfrak{R}_N(\mathcal{F}^{\text{diag}}) + \eta(\gamma)^d \mathfrak{R}_N(\mathcal{F}^{\text{coh}}).$$

Logo,

$$R(\hat{\theta}_\gamma, \gamma) \leq \varepsilon_{\text{train}} + C_{\text{fit}}(1 - \eta(\gamma)^d) + 2\mathfrak{R}_N(\mathcal{F}^{\text{diag}}) + 2\eta(\gamma)^d \mathfrak{R}_N(\mathcal{F}^{\text{coh}}) + 3\sqrt{\frac{\log(2/\delta)}{2N}}.$$

Passo 3: comparar com o bound em $\gamma = 0$. Em $\gamma = 0$, $\eta(0) = 1$, então

$$\mathfrak{R}_N(\mathcal{F}_0) \leq \mathfrak{R}_N(\mathcal{F}^{\text{diag}}) + \mathfrak{R}_N(\mathcal{F}^{\text{coh}}).$$

A diferença entre os bounds (para o mesmo termo de confiança) é dominada por:

$$\Delta(\gamma) := C_{\text{fit}}(1 - \eta(\gamma)^d) \; - \; 2(1 - \eta(\gamma)^d)\mathfrak{R}_N(\mathcal{F}^{\text{coh}}).$$

Assim, se

$$2\mathfrak{R}_N(\mathcal{F}^{\text{coh}}) \; > \; C_{\text{fit}},$$

então para todo $\gamma > 0$ com $\eta(\gamma)^d < 1$ (i.e. qualquer supressão de coerência) obtemos $\Delta(\gamma) < 0$, logo o bound para $R(\hat{\theta}_\gamma, \gamma)$ fica estritamente menor do que o bound para $R(\hat{\theta}_0, 0)$. Em particular, existe $\gamma_1 > 0$ pequeno tal que

$$R(\hat{\theta}_{\gamma_1}, \gamma_1) < R(\hat{\theta}_0, 0).$$

Passo 4: existência de minimizador interno. Pela Hipótese (3), para $\gamma \geq \bar{\gamma}$, $R(\hat{\theta}_\gamma, \gamma) \geq R_{\text{chance}} - \epsilon$, isto é, a função de desempenho degrada para ruído alto. Pelo Lema 11, $\gamma \mapsto R(\hat{\theta}_\gamma, \gamma)$ é contínua (localmente). Como já mostramos que existe $\gamma_1 \in (0, \bar{\gamma})$ com risco estritamente menor do que em $\gamma = 0$, segue que o infimum em $[0, \bar{\gamma}]$ é atingido em algum $\gamma^* \in (0, \bar{\gamma})$ e satisfaz a desigualdade estrita. □

Discussão/Observação 15 (Como ler o Teorema 14 em linguagem de banca). O termo $2\mathfrak{R}_N(\mathcal{F}^{\text{coh}})$ mede a "capacidade de memorização" via coerência. A condição $2\mathfrak{R}_N(\mathcal{F}^{\text{coh}}) > C_{\text{fit}}$ formaliza: "o prejuízo de perder ajuste (treino) é menor do que o ganho de reduzir a capacidade espúria". Isto é precisamente o mecanismo de regularização implícita.

6 Robustez ao ruído e a perturbações: contração CPTP

Lema 16 (Contração da distância de traço)**.** *Para qualquer canal CPTP $\mathcal{E}$ e quaisquer estados ρ, σ,*

$$\|\mathcal{E}(\rho) - \mathcal{E}(\sigma)\|_1 \leq \|\rho - \sigma\|_1 .$$

Proof. Resultado padrão de teoria de informação quântica (monotonicidade da distância de traço sob CPTP). □

Proposição 17 (Robustez do logit sob perturbação de entrada)**.** *Se $\|M\|_\infty \leq 1$ e ℓ é L_ℓ-Lipschitz no primeiro argumento, então para quaisquer x, x':*

$$\left|\ell(f_{\theta,\gamma}(x), y) - \ell(f_{\theta,\gamma}(x'), y)\right| \leq L_\ell \left\|\rho(x) - \rho(x')\right\|_1 .$$

Além disso, para qualquer $\gamma' > \gamma$ (mais ruído, canal composto), a sensibilidade não aumenta.

Proof. Primeiro, pelo dual de Holder:

$$|f_{\theta,\gamma}(x) - f_{\theta,\gamma}(x')| = \left|\mathrm{Tr}\big(M(\sigma_{\theta,\gamma}(x) - \sigma_{\theta,\gamma}(x'))\big)\right| \leq \|M\|_\infty \left\|\sigma_{\theta,\gamma}(x) - \sigma_{\theta,\gamma}(x')\right\|_1 \leq \left\|\rho(x) - \rho(x')\right\|_1 ,$$

onde usamos: (i) unitariedade de $U(\theta)$ preserva norma-1; (ii) Lema 16 para $\mathcal{E}_\gamma^{\circ d}$. Aplicando Lipschitz de ℓ obtém-se a primeira desigualdade. A monotonicidade em γ segue porque compor com mais ruído é compor com um CPTP adicional, que não aumenta $\|\cdot\|_1$. □

Discussão/Observação 18. Esta proposição não diz que "mais ruído sempre melhora", mas diz que mais ruído tende a *suavizar* (reduzir sensibilidade), um ingrediente de robustez. O Teorema 14 mostra quando essa suavização gera *benefício estatístico* (redução do risco).

7 Convergência do treino sob objetivo suavizado (condição PL)

Agora formalizamos uma propriedade de *convergência* que costuma ser aceita em bancas: assumimos um regime onde o objetivo satisfaz uma desigualdade PL (hipótese padrão em análise de não-convexidade benigna).

Hipótese 19 (PL e suavidade do objetivo)**.** Considere o risco empírico $F_\gamma(\theta) := \hat{R}_S(\theta, \gamma)$. Assuma que, em uma região relevante do treino, F_γ é diferenciável, tem gradiente L_γ-Lipschitz e satisfaz PL:

$$\frac{1}{2}\|\nabla F_\gamma(\theta)\|^2 \geq \mu_\gamma \, (F_\gamma(\theta) - F_\gamma^\star),$$

onde $F_\gamma^\star = \inf_\theta F_\gamma(\theta)$ e $\mu_\gamma > 0$.

Teorema 20 (Convergência linear sob PL)**.** *Sob a Hipótese 19, a descida de gradiente $\theta_{t+1} = \theta_t - \alpha \nabla F_\gamma(\theta_t)$ com passo $\alpha \in (0, 1/L_\gamma]$ satisfaz*

$$F_\gamma(\theta_t) - F_\gamma^\star \leq (1 - \alpha\mu_\gamma)^t \, (F_\gamma(\theta_0) - F_\gamma^\star).$$

Proof. Prova padrão: usar desigualdade de descida por suavidade L_γ e depois aplicar PL para converter norma do gradiente em gap de função. □

Discussão/Observação 21 (Onde o ruído entra na taxa). Em muitos modelos, aumentar γ reduz L_γ (suaviza paisagem) por contração do canal e atenuação de amplitudes (especialmente coerentes), o que pode permitir passo maior e/ou taxa melhor. Entretanto, μ_γ pode também diminuir; o ganho final é um trade-off. Este bloco trata *convergência*; o Teorema 14 trata *generalização*.

8 Contraprova: refutando "ruído sempre prejudica" e limitações de modelo alternativo

8.1 Hipótese a ser refutada

Hipótese 22 (Hipótese alternativa H0: "ruído nunca ajuda")**.** Para todo procedimento de treino e todo conjunto de dados, vale:

$$\forall \gamma > 0 : \quad R(\hat{\theta}_\gamma, \gamma) \geq R(\hat{\theta}_0, 0).$$

Proposição 23 (Contraprova: existe um problema onde ruído melhora)**.** *A Hipótese 22 é falsa. Em particular, existe uma distribuição $\mathcal{D}$, um VQC/família $\mathcal{F}_\gamma$ e um procedimento de treino tal que:*

$$\exists \gamma > 0 : \quad R(\hat{\theta}_\gamma, \gamma) < R(\hat{\theta}_0, 0).$$

Prova (construtiva, passo a passo). A construção abstrai o VQC em duas componentes (diagonal e coerente), coerente sendo espúria:

Passo 1: Defina o "sinal verdadeiro" diagonal. Considere um problema em que $y \in \{-1, +1\}$ depende de uma característica $z(x) \in [-1, 1]$ (capturada pela parte diagonal), com margem:

$$\mathbb{E}[y\, z(x)] = m > 0.$$

Passo 2: Defina uma característica coerente espúria. Defina $s(x)$ tal que $\mathbb{E}[y\, s(x)] = 0$ (zero em população), mas em amostra finita

$$\hat{c}_S := \frac{1}{N} \sum_{i=1}^{N} y_i s(x_i)$$

tipicamente não é zero; por concentração, $|\hat{c}_S| = \Theta(N^{-1/2})$ com probabilidade constante. Isto é precisamente a Hipótese 8.

Passo 3: Classe de hipóteses sobreparametrizada explora s para memorizar. Considere um procedimento de treino que escolhe $\hat{\theta}_0$ maximizando o alinhamento empírico total:

$$\hat{\theta}_0 \in \arg\max_\theta \frac{1}{N} \sum_{i=1}^{N} y_i \Big(z_\theta(x_i) + s_\theta(x_i) \Big),$$

onde a parte coerente s_θ pode ser amplificada (overparameterization) até dominar z no conjunto S, mas permanece *sem correlação em população*. Isso produz um classificador com alto ajuste ao treino e risco populacional pior, pois a decisão fica contaminada por s.

Passo 4: Insira dephasing que suprime apenas o termo coerente. Pelo Lema 10, em nível $\gamma > 0$:

$$s_{\theta,\gamma}(x) = \eta(\gamma)^d \, s_{\theta,0}(x), \qquad \eta(\gamma)^d < 1,$$

reduzindo a capacidade de amplificar a parte espúria no logit final. O componente diagonal permanece mais preservado.

Passo 5: Conclusão por separação de termos. Como s é espúrio (nulo em população), sua contribuição esperada para classificação correta é zero; mas sua variância aumenta o risco. Reduzir sua amplitude por $\eta(\gamma)^d$ reduz esse termo de variância sem destruir o termo de margem m, desde que γ seja moderado. Portanto, existe $\gamma > 0$ com risco estritamente menor do que em $\gamma = 0$. □

Discussão/Observação 24. A contraprova acima é propositalmente "minimalista": ela isola o mecanismo matemático que invalida H0. O Teorema 14 fornece condições quantitativas para que a melhoria seja garantida por bounds de generalização.

8.2 Limitações: quando o benefício não pode existir

Proposição 25 (Regime sem componente espúria ou amostra infinita $\Rightarrow \gamma^* = 0$)**.** *Suponha que (i)* $f^{\mathrm{coh}}_{\theta,0}(x) \equiv 0$ *para todo* θ *relevante (ausência de coerência explorável), ou que (ii)* $N \to \infty$ *e o procedimento de treino é consistente e converge ao minimizador populacional do modelo sem ruído. Então não há benefício estrito: qualquer* $\gamma > 0$ *não melhora o risco em relação a* $\gamma = 0$*; logo* $\gamma^* = 0$*.*

Proof. (i) Se $f^{\mathrm{coh}}_{\theta,0} \equiv 0$, então o ruído não tem termo espúrio a suprimir; resta apenas degradar o sinal (direta ou indiretamente via dinâmica), não havendo mecanismo de regularização direcionada. (ii) Se $N \to \infty$ e o treino sem ruído converge ao risco populacional mínimo do modelo ideal, não existe "gap" de generalização a ser reduzido via regularização; o ruído apenas aplica um canal (processamento de informação) antes da medição e não pode criar informação nova sobre y. Logo não ocorre melhoria estrita. □

9 Discussão metodológica e implicações

Discussão/Observação 26 (O que a banca deve checar). Para transformar este manuscrito em prova vinculada ao seu experimento, a banca tipicamente exigirá:

1. Evidência de Hipótese 8: medir empiricamente que a parte coerente tem correlação instável com y (nula em validação cruzada/múltiplos seeds) mas aparece em treino.
2. Evidência de Hipótese 12: mostrar que ao aumentar γ o termo coerente do logit decai aproximadamente como $\eta(\gamma)^d$.
3. Evidência do trade-off: (a) redução do gap (complexidade) vs. (b) aumento do erro empírico para γ grande.

Discussão/Observação 27 (Relação com literatura). Barren plateaus e trainability em VQAs são amplamente discutidos (McClean et al.; Wang et al.; Cerezo et al.). A generalização em QML sob amostra finita e capacidade efetiva também é tema ativo (Caro et al.). Este manuscrito não depende desses resultados, mas é compatível: ruído pode (i) prejudicar trainability em certos regimes, e (ii) atuar como regularização em outros — o ponto central é a condição de coerência espúria + sobreparametrização + amostra finita.

10 Conclusão

Provamos um teorema de existência de ruído benéfico condicionado e apresentamos resultados de robustez e convergência com hipóteses matematicamente claras. O fenômeno é intrinsecamente *estatístico*: surge quando o modelo tem capacidade de explorar coerências que aparecem no treino por flutuação amostral, mas não persistem na população. O ruído (especialmente dephasing) suprime justamente essas coerências, reduzindo a complexidade efetiva e melhorando generalização. Em regimes sem componente espúria (ou com amostra infinita), o benefício desaparece.

Referências (com DOI)

- McCLEAN, J. R. et al. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9, 4812, 2018. DOI: 10.1038/s41467-018-07090-4.

- WANG, S. et al. Noise-induced barren plateaus in variational quantum algorithms. *Nature Communications*, 12, 6961, 2021. DOI: 10.1038/s41467-021-27045-6.
- CEREZO, M. et al. Variational quantum algorithms. *Nature Reviews Physics*, 3, p. 625–644, 2021. DOI: 10.1038/s42254-021-00348-9.
- PRESKILL, J. Quantum computing in the NISQ era and beyond. *Quantum*, 2, 79, 2018. DOI: 10.22331/q-2018-08-06-79.
- CARO, M. C. et al. Generalization in quantum machine learning from few training data. *Nature Communications*, 13, 4919, 2022. DOI: 10.1038/s41467-022-32550-3.