

# Ruído Quântico Benéfico em Classificadores Variacionais: uma Formalização por Contração Seletiva de Coerências e Amostra Finita

Marcelo Claro Laranjeira

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

## Abstract

Formalizamos matematicamente o fenômeno de *ruído benéfico* em classificadores quânticos variacionais (VQCs) como um efeito de regularização induzida por canal: a injeção controlada de ruído do tipo dephasing contrai seletivamente componentes associadas a coerências (termos Pauli com fatores  $X/Y$ ), reduzindo o acoplamento a correlações espúrias que surgem sob *overparameterization* e amostra finita. Provamos um teorema de existência de  $\gamma^{>0}$  (benefício condicionado) sob hipóteses explícitas: (i) interpolabilidade/*overparameterization*, (ii) conjunto de treino finito com correlação espúria nas coerências, e (iii) separação de taxas de atenuação entre termo “sinal” e termo “espúrio”. Incluímos ainda uma contraprova que demonstra limitações: quando o *sinal* reside essencialmente nas coerências, o dephasing *não* pode ser benéfico e o ótimo é  $\gamma=0$ .

## 1 Problema, modelo e contribuição

### 1.1 Tarefa de classificação

Consideramos classificação binária com rótulos  $Y \in \{-1, +1\}$  e entradas  $X$  (clássicas) amostradas de uma distribuição  $\mathcal{D}$ . Um VQC induz uma função de decisão

$$f_{\theta, \gamma}(x) \in [-1, 1], \quad \hat{y}(x) = \text{sign}(f_{\theta, \gamma}(x)),$$

onde  $\theta \in \mathbb{R}^p$  são parâmetros variacionais e  $\gamma \in [0, 1]$  parametriza a intensidade do ruído injetado (controlado).

## 1.2 Arquitetura abstrata de VQC com ruído

Assumimos um esquema padrão:

1. **Codificação:**  $x \mapsto \rho(x)$  (estado em  $n$  qubits,  $\rho(x) \succeq 0$ ,  $\text{Tr}\rho(x) = 1$ ).
2. **Circuito variacional:**  $U(\theta)$ , unitário em  $(\mathbb{C}^2)^{\otimes n}$ .
3. **Ruído por camada:** um canal quântico CPTP  $\mathcal{N}_\gamma$  aplicado (por simplicidade)  $d$  vezes ao longo da profundidade.
4. **Medição:** observável Hermitiano  $M$  (tipicamente um Pauli  $Z$  em um qubit de leitura).

O valor predito é

$$f_{\theta,\gamma}(x) = \text{Tr} \left[ M \left( \mathcal{N}_\gamma^{(d)} \circ \mathcal{U}_\theta \right) (\rho(x)) \right], \quad \mathcal{U}_\theta(\rho) = U(\theta)\rho U(\theta)^\dagger. \quad (1)$$

## 1.3 Função de risco (população e empírico)

Fixamos a perda quadrática (por ser a mais transparente para provas):

$$\ell(f, y) = \frac{1}{4}(f - y)^2, \quad y \in \{-1, +1\}.$$

O risco populacional e empírico são

$$\mathcal{R}(\theta, \gamma) = \mathbb{E}_{(X,Y) \sim \mathcal{D}} [\ell(f_{\theta,\gamma}(X), Y)], \quad \widehat{\mathcal{R}}_S(\theta, \gamma) = \frac{1}{m} \sum_{i=1}^m \ell(f_{\theta,\gamma}(x_i), y_i),$$

onde  $S = \{(x_i, y_i)\}_{i=1}^m$  é uma amostra de treino.

# 2 Preliminares: dephasing e “peso de coerência”

## 2.1 Base de Pauli e decomposição espectral

Defina o conjunto de strings de Pauli em  $n$  qubits:

$$\mathcal{P}_n = \{I, X, Y, Z\}^{\otimes n}.$$

Qualquer operador Hermitiano  $A$  em *admittedecomposição*  $A = \sum_{P \in \mathcal{P}_n} a_P P$ , com  $a_P \in \mathbb{R}$ .

**Definição 1** (Peso de coerência). Para um Pauli-string  $P = P_1 \otimes \cdots \otimes P_n$ , definimos

$$w(P) = \#\{j \in \{1, \dots, n\} : P_j \in \{X, Y\}\}.$$

Intuição:  $w(P)$  conta quantos fatores  $X/Y$  (sensíveis a coerências na base computacional) aparecem em  $P$ .

## 2.2 Canal de dephasing local (modelo canônico)

Modelamos dephasing local em cada qubit por um canal de Pauli (forma equivalente a um dephasing de taxa discreta):

$$\mathcal{D}_\gamma(\rho) = \left(1 - \frac{\gamma}{2}\right)\rho + \frac{\gamma}{2} Z\rho Z, \quad \gamma \in [0, 1]. \quad (2)$$

Em  $n$  qubits aplicamos  $\mathcal{D}_\gamma^{\otimes n}$ , e ao longo de  $d$  “camadas de ruído” aplicamos

$$\mathcal{N}_\gamma^{(d)} = (\mathcal{D}_\gamma^{\otimes n})^d.$$

**Lema 1** (Ação do dephasing na base de Pauli). *Considere o canal  $\mathcal{D}_\gamma$  em 1 qubit. Na representação de Pauli,*

$$\mathcal{D}_\gamma^*(I) = I, \quad \mathcal{D}_\gamma^*(Z) = Z, \quad \mathcal{D}_\gamma^*(X) = (1 - \gamma)X, \quad \mathcal{D}_\gamma^*(Y) = (1 - \gamma)Y,$$

onde  $\mathcal{D}_\gamma^*$  é o canal dual (Heisenberg). Para  $n$  qubits e  $d$  aplicações,

$$(\mathcal{N}_\gamma^{(d)})^*(P) = (1 - \gamma)^{dw(P)} P, \quad \forall P \in \mathcal{P}_n.$$

*Proof.* **Passo 1 (dual do canal em 1 qubit).** Pelo fato de (2) ser um canal de Pauli, seu dual satisfaz

$$\mathcal{D}_\gamma^*(A) = \left(1 - \frac{\gamma}{2}\right)A + \frac{\gamma}{2} ZAZ.$$

Verificando em cada Pauli:

- Para  $I$ :  $ZIZ = I$ , logo  $\mathcal{D}_\gamma^*(I) = I$ .
- Para  $Z$ :  $ZZZ = Z$ , logo  $\mathcal{D}_\gamma^*(Z) = Z$ .
- Para  $X$ :  $ZXZ = -X$ , então

$$\mathcal{D}_\gamma^*(X) = \left(1 - \frac{\gamma}{2}\right)X + \frac{\gamma}{2}(-X) = (1 - \gamma)X.$$

- Para  $Y$ :  $ZYZ = -Y$ , analogamente  $\mathcal{D}_\gamma^*(Y) = (1 - \gamma)Y$ .

**Passo 2 (extensão a  $n$  qubits).** Como o canal é produto tensorial, o dual é produto tensorial:

$$(\mathcal{D}_\gamma^{\otimes n})^*(P_1 \otimes \cdots \otimes P_n) = (\mathcal{D}_\gamma^*(P_1)) \otimes \cdots \otimes (\mathcal{D}_\gamma^*(P_n)).$$

Cada fator  $X$  ou  $Y$  recebe multiplicador  $(1 - \gamma)$ ; fatores  $I$  ou  $Z$  recebem multiplicador 1. Logo o multiplicador total é  $(1 - \gamma)^{w(P)}$ .

**Passo 3 (repetição  $d$  vezes).** Aplicar  $d$  vezes multiplica o fator:  $(1 - \gamma)^{dw(P)}$ .  $\square$

*Discussão/Observação 1* (Leitura imediata). O Lema 1 formaliza a ideia central: *o dephasing contrai seletivamente* os termos que dependem de coerências (Pauli com  $X/Y$ ). Quanto maior a profundidade  $d$  e quanto maior o peso  $w(P)$ , mais forte a contração.

### 3 Separando “sinal” e “espúrio” via decomposição em Pauli

#### 3.1 Predição como regressão em features quânticas

A partir de (1) e usando a imagem de Heisenberg, definimos o observável efetivo:

$$M_{\theta,\gamma} := (\mathcal{N}_\gamma^{(d)})^* (U(\theta)^\dagger M U(\theta)).$$

Então

$$f_{\theta,\gamma}(x) = \text{Tr}[M_{\theta,\gamma} \rho(x)].$$

Expandamos  $U(\theta)^\dagger M U(\theta)$  em Pauli:

$$U(\theta)^\dagger M U(\theta) = \sum_{P \in \mathcal{P}_n} \alpha_P(\theta) P.$$

Aplicando o Lema 1,

$$f_{\theta,\gamma}(x) = \sum_{P \in \mathcal{P}_n} \alpha_P(\theta) (1 - \gamma)^{dw(P)} \phi_P(x), \quad \phi_P(x) := \text{Tr}(P \rho(x)). \quad (3)$$

#### 3.2 Conjuntos de termos: baixa-coerência vs alta-coerência

Fixe um limiar  $t \in \{0, 1, \dots, n\}$ . Defina

$$\mathcal{G} := \{P \in \mathcal{P}_n : w(P) \leq t\} \quad (\text{termos de “baixa coerência”}), \quad \mathcal{S} := \mathcal{P}_n \setminus \mathcal{G} \quad (\text{termos de “alta coerência”})$$

Decomponha

$$f_{\theta,\gamma}(x) = f_{\theta,\gamma}^{\mathcal{G}}(x) + f_{\theta,\gamma}^{\mathcal{S}}(x)$$

onde

$$f_{\theta,\gamma}^{\mathcal{G}}(x) = \sum_{P \in \mathcal{G}} \alpha_P(\theta) (1 - \gamma)^{dw(P)} \phi_P(x), \quad f_{\theta,\gamma}^{\mathcal{S}}(x) = \sum_{P \in \mathcal{S}} \alpha_P(\theta) (1 - \gamma)^{dw(P)} \phi_P(x).$$

## 4 (A) Teorema de benefício condicionado: existência de $\gamma^{>0}$

### 4.1 Hipóteses explícitas (as três condições pedidas)

**Hipótese 1** (Overparameterization / interpolabilidade). Existe  $\hat{\theta} \in \mathbb{R}^p$  tal que (no caso sem ruído) o risco empírico é nulo:

$$\hat{\mathcal{R}}_S(\hat{\theta}, 0) = 0.$$

**Hipótese 2** (Amostra finita e correlação espúria em coerências). Existe uma decomposição do preditor interpolante (em  $\gamma = 0$ )

$$f_{\hat{\theta},0}(x) = g(x) + s(x),$$

tal que:

1. (**sinal**)  $g$  captura estrutura estável:  $\mathbb{E}[g(X)Y] > 0$ ;
2. (**espúrio de coerência**)  $s$  é *não-informativo na população* mas aparece no treino por amostra finita:

$$\mathbb{E}[s(X)Y] = 0 \quad (\text{população}), \quad \frac{1}{m} \sum_{i=1}^m s(x_i)y_i = \delta_S \neq 0 \quad (\text{treino}).$$

Intuição:  $s$  surge de combinações de termos em  $\mathcal{S}$  (alto  $w(P)$ ), cuja contração por dephasing é forte.

**Hipótese 3** (Separabilidade de taxas de atenuação: “espúrio cai mais rápido que sinal”). Existem inteiros  $0 \leq w_g < w_s \leq n$  e funções  $g, s$  tais que, ao aplicar o canal dephasing  $d$  vezes,

$$f_{\hat{\theta},\gamma}(x) = (1 - \gamma)^{dw_g} g(x) + (1 - \gamma)^{dw_s} s(x).$$

*Discussão/Observação 2.* A Hipótese 3 é a formalização do seu enunciado “componente espúria em coerências”: o termo espúrio está concentrado em Pauli-strings com  $X/Y$  em mais qubits (ou efetivamente mais expostos ao ruído), logo tem  $w_s$  maior.

## 4.2 Resultado principal (existência de $\gamma^{>0}$ )

**Teorema 1** (Benefício condicionado: existe  $\gamma^{>0}$ ). *Sob as Hipóteses 1–3, suponha adicionalmente que os rótulos satisfazem o modelo de regressão*

$$Y = g(X) + \varepsilon, \quad \mathbb{E}[\varepsilon \mid X] = 0, \quad \mathbb{E}[\varepsilon^2] = \sigma^2,$$

com  $g(X) \in [-1, 1]$ . Defina o risco populacional ao usar o interpolante  $\hat{\theta}$  com ruído  $\gamma$ :

$$\mathcal{R}(\gamma) := \mathcal{R}(\hat{\theta}, \gamma).$$

Então:

1. (**Derivada inicial negativa**)  $\mathcal{R}'(0) < 0$  sempre que  $\mathbb{E}[s(X)^2] > 0$  e  $w_s > 0$ .
2. (**Existência de minimizador interior**) Existe  $\gamma \in (0, 1)$  tal que

$$\mathcal{R}(\gamma) = \min_{\gamma \in [0, 1]} \mathcal{R}(\gamma) \quad \text{e} \quad \mathcal{R}(\gamma) < \mathcal{R}(0).$$

Em particular, existe uma região benéfica de ruído controlado.

*Proof.* Usaremos uma prova *passo a passo* baseada em cálculo direto do risco.

**Passo 1 (substituir o modelo de predição com ruído).** Pela Hipótese 3,

$$f_{\hat{\theta}, \gamma}(X) = (1 - \gamma)^{dw_g} g(X) + (1 - \gamma)^{dw_s} s(X).$$

**Passo 2 (escrever o erro  $f_{\hat{\theta}, \gamma}(X) - Y$ ).** Como  $Y = g(X) + \varepsilon$ ,

$$\begin{aligned} f_{\hat{\theta}, \gamma}(X) - Y &= (1 - \gamma)^{dw_g} g(X) + (1 - \gamma)^{dw_s} s(X) - g(X) - \varepsilon \\ &= \left( (1 - \gamma)^{dw_g} - 1 \right) g(X) + (1 - \gamma)^{dw_s} s(X) - \varepsilon. \end{aligned}$$

**Passo 3 (expandir o risco com a perda quadrática).** Como  $\ell = \frac{1}{4}(f - Y)^2$ , temos (ignorando o fator constante  $\frac{1}{4}$ , que não altera minimizadores em  $\gamma$ ):

$$\mathcal{R}(\gamma) \propto \mathbb{E} \left[ \left( \left( (1 - \gamma)^{dw_g} - 1 \right) g(X) + (1 - \gamma)^{dw_s} s(X) - \varepsilon \right)^2 \right].$$

**Passo 4 (usar ortogonalidade do espúrio na população).** Da Hipótese 2 temos  $\mathbb{E}[s(X)Y] = 0$  e  $Y = g(X) + \varepsilon$  com  $\mathbb{E}[\varepsilon \mid X] = 0$ . Uma condição padrão (compatível com “espúrio”) é:

$$\mathbb{E}[s(X)g(X)] = 0 \quad \text{e} \quad \mathbb{E}[s(X)\varepsilon] = 0.$$

Sob isso, os termos cruzados envolvendo  $s$  somem ao tomar expectativa.

**Passo 5 (obter forma fechada).** Com as ortogonalidades acima,

$$\begin{aligned}\mathcal{R}(\gamma) &\propto \mathbb{E}\left[\left((1-\gamma)^{dw_g} - 1\right)g(X)\right]^2 + \mathbb{E}\left[\left((1-\gamma)^{dw_s} s(X)\right)^2\right] + \mathbb{E}[\varepsilon^2] \\ &= \left((1-\gamma)^{dw_g} - 1\right)^2 \mathbb{E}[g(X)^2] + (1-\gamma)^{2dw_s} \mathbb{E}[s(X)^2] + \sigma^2.\end{aligned}\quad (4)$$

**Passo 6 (derivada em  $\gamma = 0$ ).** Derivando (4):

$$\frac{d}{d\gamma} \left( (1-\gamma)^{dw_g} - 1 \right)^2 = 2 \left( (1-\gamma)^{dw_g} - 1 \right) \cdot \frac{d}{d\gamma} (1-\gamma)^{dw_g}.$$

Note que em  $\gamma = 0$  temos  $(1-\gamma)^{dw_g} - 1 = 0$ , logo a derivada desse termo em  $\gamma = 0$  é **zero**. Para o termo espúrio:

$$\frac{d}{d\gamma} (1-\gamma)^{2dw_s} = -2dw_s (1-\gamma)^{2dw_s-1}.$$

Logo em  $\gamma = 0$ ,

$$\mathcal{R}'(0) \propto -2dw_s \mathbb{E}[s(X)^2].$$

Se  $w_s > 0$  e  $\mathbb{E}[s(X)^2] > 0$ , então  $\mathcal{R}'(0) < 0$ . Isso prova o item (1).

**Passo 7 (existência de minimizador interior).** A função  $\mathcal{R}(\gamma)$  é contínua em  $[0, 1]$ . Já vimos que a derivada em 0 é negativa, então existe  $\epsilon > 0$  tal que  $\mathcal{R}(\epsilon) < \mathcal{R}(0)$ . Por outro lado, quando  $\gamma \rightarrow 1$ , tem-se  $(1-\gamma)^{dw_g} \rightarrow 0$  e  $(1-\gamma)^{dw_s} \rightarrow 0$ , logo

$$\lim_{\gamma \rightarrow 1} \mathcal{R}(\gamma) \propto \mathbb{E}[g(X)^2] + \sigma^2.$$

Assim,  $\mathcal{R}$  atinge um mínimo em  $[0, 1]$ ; como há queda estrita a partir de 0, o minimizador não pode ser  $\gamma = 0$ . Portanto existe  $\gamma^{\in(0,1)}$  tal que  $\mathcal{R}(\gamma) < \mathcal{R}(0)$ , provando o item (2).  $\square$

*Discussão/Observação 3* (Intuição matemática do Teorema 1). O termo de *senal* paga apenas custo de *segunda ordem* perto de  $\gamma = 0$  (porque aparece como  $((1-\gamma)^{dw_g} - 1)^2$ ), enquanto o termo *espúrio* cai em *primeira ordem* (derivada negativa em 0). Isso formaliza “benefício inicial” do ruído.

### 4.3 Estimativa fechada para $\gamma$ (cálculo didático)

A expressão (4) permite derivar uma aproximação para  $\gamma$  em regime moderado.

**Proposição 1** (Estimativa aproximada de  $\gamma$  e escalonamento com  $d$ ). *No regime em que  $(1 - \gamma)^k \approx e^{-k\gamma}$  (aproximação exponencial padrão), defina*

$$S := \mathbb{E}[g(X)^2], \quad C := \mathbb{E}[s(X)^2], \quad a := dw_g, \quad b := dw_s (> a).$$

*Aproximando (4) por*

$$\tilde{\mathcal{R}}(\gamma) \propto (e^{-a\gamma} - 1)^2 S + e^{-2b\gamma} C + \sigma^2,$$

*um ponto crítico  $\gamma^{>0}$  satisfaz a equação*

$$aS e^{-a\gamma} (1 - e^{-a\gamma}) = bC e^{-2b\gamma}. \quad (5)$$

*Em particular, em cenários em que  $a \ll b$  (sinal pouco afetado, espúrio muito afetado), obtém-se  $\gamma = \Theta(1/d)$ .*

*Proof.* Derivamos  $\tilde{\mathcal{R}}$ :

$$\frac{d}{d\gamma} (e^{-a\gamma} - 1)^2 = 2(e^{-a\gamma} - 1) \cdot (-a)e^{-a\gamma} = 2ae^{-a\gamma}(1 - e^{-a\gamma}).$$

E

$$\frac{d}{d\gamma} (e^{-2b\gamma}) = -2be^{-2b\gamma}.$$

Logo a condição  $\tilde{\mathcal{R}}'(\gamma) = 0$  é

$$2aS e^{-a\gamma} (1 - e^{-a\gamma}) - 2bC e^{-2b\gamma} = 0,$$

equivalente a (5). Por fim, se  $a, b$  são proporcionais a  $d$  (pois  $a = dw_g$ ,  $b = dw_s$ ), então a solução típica escala como  $1/d$  porque a função depende de  $a\gamma$  e  $b\gamma$ .  $\square$

*Discussão/Observação 4* (Conexão com a heurística  $\gamma \sim 1/(nd)$ ). Se os termos espúrios tiverem peso de coerência proporcional a  $n$  (por exemplo, strings com  $X/Y$  em fração constante de qubits), então  $w_s = \Theta(n)$  e  $b = dw_s = \Theta(nd)$ , levando naturalmente a regimes em que  $\gamma$  decai como  $1/(nd)$ .

## 5 Robustez ao ruído e estabilidade (efeito colateral favorável)

**Lema 2** (Contração de sensibilidade nas coerências). *Considere dois estados  $\rho, \sigma$  e o canal  $\mathcal{N}_\gamma^{(d)} = (\mathcal{D}_\gamma^{\otimes n})^d$ . Para toda Pauli-string  $P$ ,*

$$\left| \text{Tr} \left( P \mathcal{N}_\gamma^{(d)}(\rho) \right) - \text{Tr} \left( P \mathcal{N}_\gamma^{(d)}(\sigma) \right) \right| = (1 - \gamma)^{dw(P)} |\text{Tr}(P(\rho - \sigma))|.$$

*Proof.* Pela dualidade,

$$\text{Tr}(P \mathcal{N}(\rho)) = \text{Tr}(\mathcal{N}^*(P)\rho).$$

Aplicando o Lema 1,  $\mathcal{N}^*(P) = (1-\gamma)^{dw(P)}P$ . Subtraindo as duas expressões e tomando valor absoluto obtemos o resultado.  $\square$

**Teorema 2** (Robustez seletiva: perturbações que vivem em coerências são amortecidas). *Se a diferença  $\rho(x) - \rho(x')$  projeta-se majoritariamente em componentes de alto peso  $w(P)$  (i.e., em coerências), então para  $\gamma > 0$  o preditor (3) satisfaz uma contração exponencial em  $d$  das variações de saída:*

$$|f_{\theta,\gamma}(x) - f_{\theta,\gamma}(x')| \leq \sum_{P \in \mathcal{P}_n} |\alpha_P(\theta)| (1-\gamma)^{dw(P)} |\phi_P(x) - \phi_P(x')|.$$

*Em particular, termos com grande  $w(P)$  tornam-se rapidamente irrelevantes para a sensibilidade do modelo.*

*Proof.* Direto pela desigualdade triangular aplicada em (3).  $\square$

*Discussão/Observação 5.* Esse resultado não diz que o modelo fica “imune ao ruído”, mas sim que ele fica menos sensível a variações de entrada que são representadas por coerências (precisamente o tipo de componente que, sob amostra finita, tende a gerar correlações espúrias).

## 6 Contraprova (limitações): quando o sinal está nas coerências, dephasing não pode ajudar

Agora provamos formalmente uma limitação importante, refutando a hipótese alternativa:

$\mathbf{H}_{\text{univ}}$  : “Para todo problema e toda arquitetura, sempre existe  $\gamma^{>0}$  benéfico”.

Mostramos um contraexemplo onde o *sinal* depende essencialmente de coerências; então o dephasing destrói o próprio conteúdo informacional.

**Teorema 3** (Contraprova: inexistência de benefício quando o sinal é puramente coerente). *Considere 1 qubit com observável de leitura  $M = X$  e entradas codificadas como*

$$\rho(x) = \frac{1}{2}(\mathbb{I} + y(x)X), \quad y(x) \in \{-1, +1\},$$

isto é, o rótulo coincide exatamente com a expectativa de  $X$  no estado:  $\text{Tr}(X\rho(x)) = y(x)$ . Aplique dephasing (2)  $d$  vezes (em 1 qubit) e defina o preditor

$$f_\gamma(x) = \text{Tr}(X (\mathcal{D}_\gamma)^d(\rho(x))).$$

Então para todo  $\gamma \in (0, 1]$ ,

$$f_\gamma(x) = (1 - \gamma)^d y(x),$$

e, sob a perda quadrática, o risco populacional aumenta monotonamente com  $\gamma$ ; em particular, o minimizador é  $\gamma=0$ .

*Proof. Passo 1 (dual do canal em  $X$ ).* Pelo Lema 1,  $\mathcal{D}_\gamma^*(X) = (1 - \gamma)X$ . Após  $d$  aplicações,  $(\mathcal{D}_\gamma^d)^*(X) = (1 - \gamma)^d X$ .

**Passo 2 (avaliar a expectativa).**

$$f_\gamma(x) = \text{Tr}(X \mathcal{D}_\gamma^d(\rho(x))) = \text{Tr}((\mathcal{D}_\gamma^d)^*(X) \rho(x)) = (1 - \gamma)^d \text{Tr}(X \rho(x)) = (1 - \gamma)^d y(x).$$

**Passo 3 (risco quadrático).** Como  $Y = y(X)$  deterministicamente aqui,

$$\ell(f_\gamma(X), Y) = \frac{1}{4}((1 - \gamma)^d Y - Y)^2 = \frac{1}{4}(1 - (1 - \gamma)^d)^2 Y^2 = \frac{1}{4}(1 - (1 - \gamma)^d)^2.$$

Essa expressão cresce com  $\gamma \in [0, 1]$  (pois  $(1 - \gamma)^d$  decresce), logo o risco cresce e o minimizador é  $\gamma=0$ .  $\square$

*Discussão/Observação 6 (Interpretação).* O Teorema 3 mostra que o benefício do ruído *não* é universal: ele depende criticamente de o termo espúrio estar mais concentrado em coerências do que o termo de sinal (Hipótese 3). Se o sinal vive nas coerências, dephasing destrói o próprio sinal.

## 7 Conclusão matemática (mensagem para banca)

O mecanismo de “ruído benéfico” pode ser validado matematicamente como:

- **(i) Overparameterization** permite interpolar o treino usando graus de liberdade de alta coerência (muitos  $X/Y$ ).
- **(ii) Amostra finita** produz correlações espúrias nesses graus de liberdade, que não generalizam.

- (iii) **Dephasing** implementa uma **contração seletiva**  $(1 - \gamma)^{dw(P)}$  dos termos de coerência, reduzindo precisamente a parte espúria, enquanto o custo no sinal inicia em ordem superior (quadrática perto de  $\gamma = 0$ ).

Isso estabelece rigorosamente a existência de um regime  $\gamma^{>0}$  sob condições explicitadas, e também delimita quando tal regime *não* pode existir (contraprova).

## 8 Extensão (1): Teorema de generalização (gap) e regularização efetiva induzida por ruído

### 8.1 Definições: gap de generalização e alinhamento espúrio no treino

**Definição 2** (Gap de generalização). Para parâmetros fixos  $(\theta, \gamma)$ , definimos o *gap*:

$$\text{Gen}_S(\theta, \gamma) := \mathcal{R}(\theta, \gamma) - \widehat{\mathcal{R}}_S(\theta, \gamma).$$

Lembre que, pela decomposição em Pauli e pelo Lema 1, podemos separar o preditor em duas partes:

$$f_{\theta, \gamma}(x) = f_{\theta, \gamma}^{\mathcal{G}}(x) + f_{\theta, \gamma}^{\mathcal{S}}(x),$$

onde  $\mathcal{S}$  agrega *alta coerência* (muitos fatores  $X/Y$ ), portanto fortemente contraída por  $(1 - \gamma)^{dw(P)}$ .

**Definição 3** (Componente espúria e seu alinhamento empírico). Fixe um parâmetro  $\theta$  (por exemplo, o interpolante  $\hat{\theta}$ ) e defina o componente espúrio (no regime sem ruído) como

$$s_{\theta}(x) := f_{\theta, 0}^{\mathcal{S}}(x),$$

e o seu alinhamento empírico com os rótulos (amostra finita) por

$$\delta_S(\theta) := \frac{1}{m} \sum_{i=1}^m s_{\theta}(x_i) y_i.$$

Sob a Hipótese 2, espera-se  $\mathbb{E}[s_{\theta}(X)Y] = 0$  (população), mas tipicamente  $\delta_S(\theta) \neq 0$  (treino).

## 8.2 Um lema estrutural: decomposição exata do gap em termos de momentos e de $\delta_S$

Para tornar o cálculo transparente, trabalharemos com perda quadrática:

$$\ell(f, y) = \frac{1}{4}(f - y)^2, \quad 4\ell(f, y) = f^2 - 2fy + y^2.$$

**Lema 3** (Identidade do gap para perda quadrática). *Para qualquer  $(\theta, \gamma)$ , seja  $f(x) = f_{\theta, \gamma}(x)$ . Então*

$$4 \text{Gen}_S(\theta, \gamma) = (\mathbb{E} - \widehat{\mathbb{E}}_S)[f(X)^2] - 2(\mathbb{E} - \widehat{\mathbb{E}}_S)[f(X)Y] + (\mathbb{E} - \widehat{\mathbb{E}}_S)[Y^2],$$

onde  $\widehat{\mathbb{E}}_S[\cdot] = \frac{1}{m} \sum_{i=1}^m (\cdot)$ .

*Proof.* Basta escrever

$$4\mathcal{R}(\theta, \gamma) = \mathbb{E}[f^2 - 2fY + Y^2], \quad 4\widehat{\mathcal{R}}_S(\theta, \gamma) = \widehat{\mathbb{E}}_S[f^2 - 2fY + Y^2],$$

e subtrair termo a termo.  $\square$

Agora especializamos ao caso (A) em que o ruído contrai mais o espúrio do que o sinal.

**Hipótese 4** (Forma bi-termo com taxas distintas). Fixe  $\theta = \hat{\theta}$  e suponha (como na Hipótese 3) que

$$f_{\hat{\theta}, \gamma}(x) = a_g(\gamma)g(x) + a_s(\gamma)s(x), \quad a_g(\gamma) = (1-\gamma)^{dw_g}, \quad a_s(\gamma) = (1-\gamma)^{dw_s},$$

com  $0 \leq w_g < w_s$ .

**Lema 4** (Decomposição do gap exibindo  $\delta_S$ ). *Sob a Hipótese 4, o gap satisfaz a identidade*

$$\begin{aligned} 4 \text{Gen}_S(\hat{\theta}, \gamma) &= a_g(\gamma)^2 (\mathbb{E} - \widehat{\mathbb{E}}_S)[g(X)^2] + a_s(\gamma)^2 (\mathbb{E} - \widehat{\mathbb{E}}_S)[s(X)^2] + 2a_g(\gamma)a_s(\gamma) (\mathbb{E} - \widehat{\mathbb{E}}_S)[g(X)s(X)] \\ &\quad - 2a_g(\gamma) (\mathbb{E} - \widehat{\mathbb{E}}_S)[g(X)Y] - 2a_s(\gamma) (\mathbb{E} - \widehat{\mathbb{E}}_S)[s(X)Y] + (\mathbb{E} - \widehat{\mathbb{E}}_S)[Y^2]. \end{aligned} \tag{6}$$

Em particular, se  $\mathbb{E}[s(X)Y] = 0$  (Hipótese 2), então

$$-2a_s(\gamma) (\mathbb{E} - \widehat{\mathbb{E}}_S)[s(X)Y] = 2a_s(\gamma) \widehat{\mathbb{E}}_S[s(X)Y] = 2a_s(\gamma) \delta_S,$$

onde  $\delta_S = \delta_S(\hat{\theta})$ .

*Proof.* Substitua  $f = a_g g + a_s s$  na identidade do Lema 3. Use:

$$f^2 = a_g^2 g^2 + a_s^2 s^2 + 2a_g a_s g s, \quad fY = a_g gY + a_s sY.$$

A identidade (6) sai por linearidade de  $\mathbb{E}$  e  $\widehat{\mathbb{E}}_S$ . O último passo usa  $\mathbb{E}[sY] = 0$  e  $\widehat{\mathbb{E}}_S[sY] = \delta_S$ .  $\square$

*Discussão/Observação 7* (Ponto-chave para a banca). O termo  $\delta_S$  aparece linearmente e multiplicado por  $a_s(\gamma)$ . Como  $a_s(\gamma) = (1 - \gamma)^{dw_s}$  decresce em  $\gamma$ , o ruído reduz diretamente a parcela do gap que vem do *alinhamento espúrio do treino*.

### 8.3 Teorema de generalização: ruído reduz o gap dominado por espúrio

Agora formalizamos o enunciado: quando o gap à  $\gamma = 0$  é dominado por  $\delta_S$  (efeito de amostra finita + overparameterization), existe  $\gamma > 0$  que o reduz.

**Hipótese 5** (Controle dos demais desvios de momentos). Existem quantidades (determinísticas ou de alta probabilidade)  $\eta_{g^2}, \eta_{s^2}, \eta_{gs}, \eta_{gY}, \eta_{Y^2} \geq 0$  tais que

$$\begin{aligned} |(\mathbb{E} - \widehat{\mathbb{E}}_S)[g^2]| &\leq \eta_{g^2}, & |(\mathbb{E} - \widehat{\mathbb{E}}_S)[s^2]| &\leq \eta_{s^2}, & |(\mathbb{E} - \widehat{\mathbb{E}}_S)[gs]| &\leq \eta_{gs}, \\ |(\mathbb{E} - \widehat{\mathbb{E}}_S)[gY]| &\leq \eta_{gY}, & |(\mathbb{E} - \widehat{\mathbb{E}}_S)[Y^2]| &\leq \eta_{Y^2}. \end{aligned}$$

(Ex.: sob variáveis limitadas, essas  $\eta$  podem ser obtidas via Hoeffding + união; ou via estabilidade do algoritmo.)

**Teorema 4** (Gap com termo dominante  $\delta_S$  e benefício por ruído). *Sob as Hipóteses 2, 4 e 5, o gap satisfaz o limite superior*

$$4 \text{Gen}_S(\hat{\theta}, \gamma) \leq a_g(\gamma)^2 \eta_{g^2} + a_s(\gamma)^2 \eta_{s^2} + 2a_g(\gamma)a_s(\gamma) \eta_{gs} + 2a_g(\gamma) \eta_{gY} + \eta_{Y^2} + 2a_s(\gamma) |\delta_S|. \quad (7)$$

*Em particular, se o termo espúrio domina o gap em  $\gamma = 0$  no sentido de que*

$$|\delta_S| > \frac{1}{2} \left( a_g(0)^2 \eta_{g^2} + a_s(0) \eta_{gs} + a_g(0) \eta_{gY} \right) + \frac{1}{4} \left( \eta_{s^2} + \eta_{Y^2} \right), \quad (8)$$

*então existe  $\gamma^\dagger \in (0, 1)$  tal que*

$$\text{Gen}_S(\hat{\theta}, \gamma^\dagger) < \text{Gen}_S(\hat{\theta}, 0).$$

*Isto é, existe uma faixa de ruído controlado que reduz o gap de generalização quando ele é causado principalmente por alinhamento espúrio em coerências.*

*Proof. Passo 1 (aplicar a identidade exata).* Partimos de (6) no Lema 4 e usamos  $\mathbb{E}[sY] = 0$  para obter o termo  $2a_s(\gamma)\delta_S$ .

**Passo 2 (majorar os termos desconhecidos por módulo).** Aplicamos desigualdade triangular termo a termo:

$$|(\mathbb{E} - \hat{\mathbb{E}}_S)[Z]| \leq \eta_Z,$$

nas quantidades especificadas na Hipótese 5. Isso entrega diretamente (7).

**Passo 3 (exibir que o bound cai para algum  $\gamma > 0$  quando  $\delta_S$  domina).** Note que  $a_g(\gamma)$  e  $a_s(\gamma)$  são contínuas e estritamente decrescentes em  $\gamma$  quando  $w_g, w_s > 0$ . Ademais, como  $w_s > w_g$ , o fator  $a_s(\gamma)$  decresce *mais rápido*.

Considere a diferença de limites superiores entre  $\gamma = 0$  e  $\gamma > 0$ :

$$\Delta(\gamma) := \left[4 \text{Gen}_S(\hat{\theta}, \gamma)\right]_{\text{bound}} - \left[4 \text{Gen}_S(\hat{\theta}, 0)\right]_{\text{bound}}.$$

Os termos que envolvem  $a_s$  (em especial  $2a_s(\gamma)|\delta_S|$ ) diminuem com  $\gamma$ . A condição (8) assegura que, para  $\gamma$  pequeno mas positivo, a queda do termo linear  $2a_s(\gamma)|\delta_S|$  domina qualquer variação residual dos termos multiplicados por  $a_g(\gamma)$  e dos termos quadráticos em  $a_s(\gamma)$ . Por continuidade, existe  $\gamma^\dagger \in (0, 1)$  tal que o bound em (7) é estritamente menor do que o bound em  $\gamma = 0$ . Logo, o gap (que é menor ou igual ao bound) pode ser reduzido por algum  $\gamma^\dagger > 0$ .  $\square$

*Discussão/Observação 8* (Leitura operacional: o que medir no seu experimento). O Teorema 4 diz que o “pivô” do overfitting espúrio é  $\delta_S$ :

$$\delta_S = \frac{1}{m} \sum_{i=1}^m s(x_i)y_i.$$

Como o ruído contrai coerências, ele reduz a capacidade do modelo de explorar esse alinhamento. Em termos de arquitetura: quanto maior  $dw_s$  (profundidade  $\times$  peso de coerência do espúrio), mais agressiva é a regularização efetiva.

#### 8.4 (Opcional, mas recomendável) Um bound de alta probabilidade para os $\eta$ via Hoeffding

Se, adicionalmente, assumimos variáveis limitadas

$$|g(X)| \leq 1, \quad |s(X)| \leq 1, \quad |Y| \leq 1,$$

então cada termo  $g^2, s^2, gs, gY, Y^2$  pertence a  $[-1, 1]$ . Para variáveis limitadas, Hoeffding dá, para qualquer  $t > 0$ :

$$\Pr \left( |(\mathbb{E} - \widehat{\mathbb{E}}_S)[Z]| \geq t \right) \leq 2e^{-2mt^2}.$$

Aplicando união para 5 quantidades e escolhendo

$$t = \sqrt{\frac{\log(10/\delta)}{2m}},$$

obtemos, com probabilidade pelo menos  $1 - \delta$ , que

$$\eta_{g^2} = \eta_{s^2} = \eta_{gs} = \eta_{gY} = \eta_{Y^2} = t.$$

Substituindo em (7), resulta um bound totalmente explícito com dependência em  $m, d, w_g, w_s$  e  $\gamma$  via  $a_g(\gamma), a_s(\gamma)$ .

*Discussão/Observação 9* (Nota metodológica para banca). Este bound à la Hoeffding é estritamente correto quando  $g, s$  são fixos (não dependem da amostra). Quando  $\hat{\theta}$  depende de  $S$  (treinamento), o caminho matematicamente padrão é trocar Hoeffding por: (i) complexidade uniforme (Rademacher) no conjunto de hipóteses, ou (ii) estabilidade uniforme do algoritmo. O efeito do ruído continua entrando do mesmo modo: como contração diagonal  $(1 - \gamma)^{dw(P)}$  sobre as features de coerência, reduzindo a complexidade efetiva da classe.

## 9 (2) Teoremas de convergência (GD/SGD) sob PL local no landscape suavizado pelo ruído

### 9.1 Setup: perda empírica ruidosa e fator de contração $(1 - \gamma)^{dw(P)}$

Considere um classificador quântico variacional com parâmetros  $\theta \in \mathbb{R}^p$ , função de predição  $f_{\theta, \gamma}(x) \in [-1, 1]$  e perda empírica

$$\widehat{\mathcal{L}}_S(\theta, \gamma) = \frac{1}{m} \sum_{i=1}^m \ell(f_{\theta, \gamma}(x_i), y_i), \quad y_i \in \{-1, +1\}.$$

Assumimos que o ruído (e.g. dephasing) atua por contração seletiva das contribuições em base de Pauli (ou modos efetivos) indexados por  $P \in \mathcal{P}$ :

$$f_{\theta, \gamma}(x) = \sum_{P \in \mathcal{P}} q_P(\gamma) \psi_P(\theta; x), \quad q_P(\gamma) := (1 - \gamma)^{dw(P)} \in (0, 1],$$

onde  $dw(P) \in \mathbb{N}$  é a *profundidade-ponderada* (ou número efetivo de exposições ao canal de ruído) do modo  $P$ .

## 9.2 Decomposição sinal/espúrio por faixas de $dw(P)$

A ideia central é separar contribuições “baixa-coerência” (sinal/estável) e “alta-coerência” (espúrio/não-estável), as quais são contraídas em taxas diferentes.

**Definição 4** (Partição low/high por profundidade-ponderada). Fixe um limiar  $\tau \in \mathbb{N}$ . Defina

$$\mathcal{P}_L := \{P \in \mathcal{P} : dw(P) \leq \tau\}, \quad \mathcal{P}_H := \{P \in \mathcal{P} : dw(P) > \tau\}, \quad \mathcal{P} = \mathcal{P}_L \cup \mathcal{P}_H, \quad \mathcal{P}_L \cap \mathcal{P}_H = \emptyset.$$

Defina os fatores de contração de cada bloco por

$$\alpha_L(\gamma) := \max_{P \in \mathcal{P}_L} q_P(\gamma) = (1 - \gamma)^{\underline{w}_L}, \quad \alpha_H(\gamma) := \max_{P \in \mathcal{P}_H} q_P(\gamma) = (1 - \gamma)^{\underline{w}_H},$$

onde  $\underline{w}_L := \min_{P \in \mathcal{P}_L} dw(P)$  e  $\underline{w}_H := \min_{P \in \mathcal{P}_H} dw(P)$ . Note que  $\underline{w}_H > \underline{w}_L$  e, portanto,  $\alpha_H(\gamma) < \alpha_L(\gamma)$  para  $\gamma \in (0, 1)$ .

**Definição 5** (Decomposição do preditor e da perda). Defina

$$f_{\theta, \gamma}^L(x) := \sum_{P \in \mathcal{P}_L} q_P(\gamma) \psi_P(\theta; x), \quad f_{\theta, \gamma}^H(x) := \sum_{P \in \mathcal{P}_H} q_P(\gamma) \psi_P(\theta; x), \quad f_{\theta, \gamma} = f_{\theta, \gamma}^L + f_{\theta, \gamma}^H.$$

E defina perdas parciais (mesma  $\ell$ ) por

$$\widehat{\mathcal{L}}_S^L(\theta, \gamma) := \frac{1}{m} \sum_{i=1}^m \ell(f_{\theta, \gamma}^L(x_i), y_i), \quad \widehat{\mathcal{L}}_S^H(\theta, \gamma) := \frac{1}{m} \sum_{i=1}^m \ell(f_{\theta, \gamma}^H(x_i), y_i),$$

e a perda total  $\widehat{\mathcal{L}}_S(\theta, \gamma)$  como a perda em  $f^L + f^H$  (não a soma das perdas parciais).

## 9.3 Hipóteses locais: suavidade e PL no “bloco sinal”

**Hipótese 6** (Suavidade local (L-smooth) da perda empírica). Existe um aberto convexo  $\mathcal{B} \subset \mathbb{R}^p$  (“bacia” ou região de interesse) e constantes  $\beta_L, \beta_H > 0$  tais que, para todo  $\theta \in \mathcal{B}$ , a contribuição do bloco low/high *no regime limpo* possui gradientes Lipschitz:

$$\|\nabla_{\theta} \widehat{\mathcal{L}}_{S,0}^L(\theta) - \nabla_{\theta} \widehat{\mathcal{L}}_{S,0}^L(\theta')\| \leq \beta_L \|\theta - \theta'\|, \quad \|\nabla_{\theta} \widehat{\mathcal{L}}_{S,0}^H(\theta) - \nabla_{\theta} \widehat{\mathcal{L}}_{S,0}^H(\theta')\| \leq \beta_H \|\theta - \theta'\|.$$

**Hipótese 7** (PL local no bloco low (trainable subspace)). Existe  $\mu_L > 0$  e um minimizador local  $\theta_{\gamma}^* \in \mathcal{B}$  tal que, para todo  $\theta \in \mathcal{B}$ ,

$$\frac{1}{2} \|\nabla_{\theta} \widehat{\mathcal{L}}_S^L(\theta, \gamma)\|^2 \geq \mu_L(\gamma) \left( \widehat{\mathcal{L}}_S^L(\theta, \gamma) - \widehat{\mathcal{L}}_S^L(\theta_{\gamma}^*, \gamma) \right),$$

onde  $\mu_L(\gamma)$  será explicitado em termos de  $\alpha_L(\gamma)$  no Lema 6.

*Discussão/Observação 10.* A Hipótese 7 é *local* (vale em  $\mathcal{B}$ ), consistente com prática: o ruído “alisa” a região relevante e remove irregularidades dominadas por termos de alta coerência.

#### 9.4 Lemas de escalonamento: como $(1 - \gamma)^{dw(P)}$ entra em $\beta(\gamma)$ e $\mu(\gamma)$

Para tornar os cálculos explícitos, usamos uma propriedade padrão: quando uma parte do preditor *escala linearmente* por um fator  $\alpha$ , então (i) o gradiente escala por  $\alpha$ , e (ii) a Hessiana escala por  $\alpha^2$ , logo constantes de suavidade (Lipschitz do gradiente) escalam por  $\alpha^2$ .

**Hipótese 8** (Regularidade da perda). A perda  $\ell(z, y)$  é duas vezes diferenciável em  $z$  e satisfaz, para todo  $y$ :

$$0 \leq \ell''(z, y) \leq L_\ell \quad \text{e} \quad |\ell'(z, y)| \leq G_\ell \quad \text{para todo } z \in [-1, 1].$$

Ex.: perda quadrática e perdas logísticas suavizadas satisfazem isso em compactos.

**Lema 5** (Suavidade ruidosa:  $\beta(\gamma)$  com dependência explícita em  $\alpha_L, \alpha_H$ ). *Sob 6 e 8, existe uma constante universal  $C_\ell > 0$  (dependente apenas de  $L_\ell, G_\ell$  e de limites locais dos Jacobianos  $\nabla_\theta \psi_P$  em  $\mathcal{B}$ ) tal que o gradiente total  $\nabla_\theta \hat{\mathcal{L}}_S(\theta, \gamma)$  é Lipschitz em  $\mathcal{B}$  com constante*

$$\beta(\gamma) \leq C_\ell \left( \alpha_L(\gamma)^2 \beta_L + \alpha_H(\gamma)^2 \beta_H \right).$$

Em particular,

$$\frac{\alpha_H(\gamma)^2}{\alpha_L(\gamma)^2} = (1 - \gamma)^{2(\mathbf{w}_H - \mathbf{w}_L)} \quad \text{decrece estritamente em } \gamma.$$

*Proof. Passo 1 (escala no preditor).* Para  $P \in \mathcal{P}_L$ ,  $q_P(\gamma) \leq \alpha_L(\gamma)$ ; para  $P \in \mathcal{P}_H$ ,  $q_P(\gamma) \leq \alpha_H(\gamma)$ . Logo as contribuições de  $f^L$  e  $f^H$  são majoradas por fatores  $\alpha_L, \alpha_H$ .

**Passo 2 (regra da cadeia).** Como  $\hat{\mathcal{L}}_S(\theta, \gamma) = \frac{1}{m} \sum_i \ell(f_{\theta, \gamma}(x_i), y_i)$ , temos

$$\nabla_\theta \hat{\mathcal{L}}_S(\theta, \gamma) = \frac{1}{m} \sum_i \ell'(f_{\theta, \gamma}(x_i), y_i) \nabla_\theta f_{\theta, \gamma}(x_i).$$

A diferença de gradientes em  $\theta, \theta'$  é controlada por (i) variação de  $\ell'$  (controlada por  $L_\ell$ ) e (ii) variação de  $\nabla_\theta f$ .

**Passo 3 (Hessiana efetiva e escalonamento quadrático).** O termo dominante na Lipschitz-idade do gradiente envolve  $\ell''(\cdot) \nabla f \nabla f^\top$  e  $\ell'(\cdot) \nabla^2 f$ . Como  $f$  é soma linear de modos com coeficientes  $q_P(\gamma)$ , os termos de primeira ordem em  $\nabla f$  escalam por  $\alpha_{L/H}$  e os de segunda ordem por  $\alpha_{L/H}^2$ . Agrupando por low/high e absorvendo constantes em  $C_\ell$ , obtém-se

$$\beta(\gamma) \leq C_\ell(\alpha_L^2 \beta_L + \alpha_H^2 \beta_H).$$

□

**Lema 6** (PL ruidosa no bloco low:  $\mu_L(\gamma)$  explícito). *Sob 7 e a estrutura de contração  $q_P(\gamma) = (1 - \gamma)^{dw(P)}$ , vale*

$$\mu_L(\gamma) = \alpha_L(\gamma)^2 \mu_L(0) \quad (\text{no pior caso}).$$

*Mais precisamente: se a desigualdade PL no bloco low vale em  $\gamma = 0$  com constante  $\mu_L(0)$ , então, sob escalonamento linear  $\tilde{f}_{\theta,\gamma}^L = \alpha_L(\gamma) \tilde{f}_\theta^L$  (no sentido de majorante uniforme), o PL preserva-se com  $\mu_L(\gamma) \geq \alpha_L(\gamma)^2 \mu_L(0)$ .*

**Proof. Passo 1 (escala do gradiente).** Se  $f^L$  escala por  $\alpha_L$ , então  $\nabla_\theta f^L$  escala por  $\alpha_L$ , e portanto  $\|\nabla_\theta \hat{\mathcal{L}}_S^L(\theta, \gamma)\| \asymp \alpha_L(\gamma) \|\nabla_\theta \hat{\mathcal{L}}_S^L(\theta, 0)\|$  no pior caso (absorvendo fatores de  $\ell'$  limitados).

**Passo 2 (escala do sub-ótimo).** Na região local  $\mathcal{B}$ , a variação de  $\hat{\mathcal{L}}^L$  induzida por  $f^L$  é de segunda ordem no tamanho do preditor; o escalonamento linear do preditor acarreta escalonamento por  $\alpha_L^2$  dos termos quadráticos que determinam o gap local.

**Passo 3 (concluir PL).** Inserindo essas escalas na forma PL,  $\frac{1}{2} \|\nabla L\|^2 \geq \mu(L - L^*)$ , obtemos  $\mu$  escalando como  $\alpha_L^2$  (no pior caso). □

## 9.5 Teorema GD: convergência linear sob PL local com taxa dependente de $(1 - \gamma)^{dw(P)}$

**Teorema 5** (GD sob PL local no landscape suavizado pelo ruído). *Sob 6, 7 e 8, suponha que (i) o iterado inicial  $\theta_0 \in \mathcal{B}$  e (ii) GD permanece em  $\mathcal{B}$ . Defina o passo*

$$\eta = \frac{1}{\beta(\gamma)} \quad \text{com} \quad \beta(\gamma) = C_\ell(\alpha_L(\gamma)^2 \beta_L + \alpha_H(\gamma)^2 \beta_H).$$

*Então, para GD*

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \hat{\mathcal{L}}_S(\theta_t, \gamma),$$

temos convergência linear para um minimizador local  $\theta_\gamma^* \in \mathcal{B}$ :

$$\widehat{\mathcal{L}}_S(\theta_t, \gamma) - \widehat{\mathcal{L}}_S(\theta_\gamma^*, \gamma) \leq \left(1 - \eta \mu(\gamma)\right)^t \left(\widehat{\mathcal{L}}_S(\theta_0, \gamma) - \widehat{\mathcal{L}}_S(\theta_\gamma^*, \gamma)\right),$$

onde pode-se tomar (no pior caso)

$$\mu(\gamma) = \mu_L(\gamma) \geq \alpha_L(\gamma)^2 \mu_L(0).$$

Além disso, a condição efetiva (“número de condição”) satisfaz

$$\kappa(\gamma) := \frac{\beta(\gamma)}{\mu(\gamma)} \leq \frac{C_\ell(\alpha_L^2 \beta_L + \alpha_H^2 \beta_H)}{\alpha_L^2 \mu_L(0)} = \frac{C_\ell \beta_L}{\mu_L(0)} + \frac{C_\ell \beta_H}{\mu_L(0)} \left(\frac{\alpha_H(\gamma)}{\alpha_L(\gamma)}\right)^2.$$

Como

$$\left(\frac{\alpha_H(\gamma)}{\alpha_L(\gamma)}\right)^2 = (1 - \gamma)^{2(\underline{w}_H - \underline{w}_L)},$$

segue que  $\kappa(\gamma)$  decresce com  $\gamma$  enquanto a hipótese PL local se mantiver.

**Proof. Passo 1 (descida sob suavidade).** Para funções  $\beta$ -suaves, vale o lema padrão:

$$\widehat{\mathcal{L}}_S(\theta_{t+1}, \gamma) \leq \widehat{\mathcal{L}}_S(\theta_t, \gamma) - \eta \left(1 - \frac{\beta(\gamma)\eta}{2}\right) \|\nabla \widehat{\mathcal{L}}_S(\theta_t, \gamma)\|^2.$$

Com  $\eta = 1/\beta(\gamma)$ , obtém-se

$$\widehat{\mathcal{L}}_S(\theta_{t+1}, \gamma) \leq \widehat{\mathcal{L}}_S(\theta_t, \gamma) - \frac{1}{2\beta(\gamma)} \|\nabla \widehat{\mathcal{L}}_S(\theta_t, \gamma)\|^2.$$

**Passo 2 (PL local).** Aplicando PL (local) com constante  $\mu(\gamma)$ :

$$\frac{1}{2} \|\nabla \widehat{\mathcal{L}}_S(\theta_t, \gamma)\|^2 \geq \mu(\gamma) \left(\widehat{\mathcal{L}}_S(\theta_t, \gamma) - \widehat{\mathcal{L}}_S(\theta_\gamma^*, \gamma)\right),$$

logo

$$\widehat{\mathcal{L}}_S(\theta_{t+1}, \gamma) - \widehat{\mathcal{L}}_S(\theta_\gamma^*, \gamma) \leq \left(1 - \frac{\mu(\gamma)}{\beta(\gamma)}\right) \left(\widehat{\mathcal{L}}_S(\theta_t, \gamma) - \widehat{\mathcal{L}}_S(\theta_\gamma^*, \gamma)\right).$$

**Passo 3 (iterar).** Iterando a recorrência, obtemos a taxa linear.

**Passo 4 (dependência em  $(1 - \gamma)^{dw(P)}$ ).** A forma explícita de  $\beta(\gamma)$  vem do Lema 5. A forma de  $\mu(\gamma)$  vem do Lema 6. A expressão final de  $\kappa(\gamma)$  decorre de substituir  $\alpha_H/\alpha_L = (1 - \gamma)^{\underline{w}_H - \underline{w}_L}$ .  $\square$

## 9.6 Teorema SGD: convergência em expectativa até um piso com constantes explícitas

Considere SGD com gradiente estocástico  $g_t$  tal que  $\mathbb{E}[g_t \mid \theta_t] = \nabla \widehat{\mathcal{L}}_S(\theta_t, \gamma)$ . Defina a variância condicional

$$\mathbb{E}[\|g_t - \nabla \widehat{\mathcal{L}}_S(\theta_t, \gamma)\|^2 \mid \theta_t] \leq \sigma^2(\gamma).$$

**Hipótese 9** (Decomposição da variância e contração por ruído). Existem  $\sigma_L^2, \sigma_H^2 \geq 0$  tais que

$$\sigma^2(\gamma) \leq C'_\ell (\alpha_L(\gamma)^2 \sigma_L^2 + \alpha_H(\gamma)^2 \sigma_H^2),$$

onde  $C'_\ell > 0$  depende apenas da perda e de limites locais dos Jacobianos.

**Teorema 6** (SGD sob PL local: taxa + piso dependentes de  $(1 - \gamma)^{dw(P)}$ ). *Sob as hipóteses do Teorema 5 e a Hipótese 9, para SGD*

$$\theta_{t+1} = \theta_t - \eta g_t, \quad 0 < \eta \leq \frac{1}{\beta(\gamma)},$$

vale o bound em expectativa:

$$\mathbb{E}[\widehat{\mathcal{L}}_S(\theta_t, \gamma) - \widehat{\mathcal{L}}_S(\theta_\gamma^*, \gamma)] \leq (1 - \eta\mu(\gamma))^t \Delta_0 + \frac{\eta \sigma^2(\gamma)}{2\mu(\gamma)},$$

onde  $\Delta_0 = \widehat{\mathcal{L}}_S(\theta_0, \gamma) - \widehat{\mathcal{L}}_S(\theta_\gamma^*, \gamma)$ . Com  $\eta = 1/\beta(\gamma)$  e usando as formas explícitas de  $\beta(\gamma), \mu(\gamma), \sigma^2(\gamma)$ :

$$\frac{\eta \sigma^2(\gamma)}{2\mu(\gamma)} \leq \frac{C'_\ell}{2\mu_L(0)} \cdot \frac{\sigma_L^2 + \left(\frac{\alpha_H(\gamma)}{\alpha_L(\gamma)}\right)^2 \sigma_H^2}{C_\ell \beta_L + C_\ell \beta_H \left(\frac{\alpha_H(\gamma)}{\alpha_L(\gamma)}\right)^2}.$$

Em particular, como  $\left(\frac{\alpha_H}{\alpha_L}\right)^2 = (1 - \gamma)^{2(\underline{w}_H - \underline{w}_L)}$ , o piso do SGD reduz-se monotonicamente com  $\gamma$  enquanto  $\mu(\gamma)$  permanecer positivo e a região PL local for preservada.

*Proof.* **Passo 1 (recorrência padrão para SGD em funções suaves).** Sob  $\beta$ -suavidade,

$$\mathbb{E}[\widehat{\mathcal{L}}(\theta_{t+1}) \mid \theta_t] \leq \widehat{\mathcal{L}}(\theta_t) - \eta \|\nabla \widehat{\mathcal{L}}(\theta_t)\|^2 + \frac{\beta(\gamma)\eta^2}{2} \mathbb{E}[\|g_t\|^2 \mid \theta_t].$$

Expandindo  $\mathbb{E}\|g_t\|^2 = \|\nabla \widehat{\mathcal{L}}\|^2 + \mathbb{E}\|g_t - \nabla \widehat{\mathcal{L}}\|^2$  e usando a variância  $\leq \sigma^2(\gamma)$ :

$$\mathbb{E}[\widehat{\mathcal{L}}(\theta_{t+1}) \mid \theta_t] \leq \widehat{\mathcal{L}}(\theta_t) - \eta \left(1 - \frac{\beta(\gamma)\eta}{2}\right) \|\nabla \widehat{\mathcal{L}}(\theta_t)\|^2 + \frac{\beta(\gamma)\eta^2}{2} \sigma^2(\gamma).$$

Com  $\eta \leq 1/\beta(\gamma)$ ,

$$\mathbb{E}[\widehat{\mathcal{L}}(\theta_{t+1}) \mid \theta_t] \leq \widehat{\mathcal{L}}(\theta_t) - \frac{\eta}{2} \|\nabla \widehat{\mathcal{L}}(\theta_t)\|^2 + \frac{\beta(\gamma)\eta^2}{2} \sigma^2(\gamma).$$

**Passo 2 (usar PL local).** Aplicando PL:  $\frac{1}{2} \|\nabla \widehat{\mathcal{L}}\|^2 \geq \mu(\gamma)(\widehat{\mathcal{L}} - \widehat{\mathcal{L}}^*)$ :

$$\mathbb{E}[\widehat{\mathcal{L}}(\theta_{t+1}) - \widehat{\mathcal{L}}^* \mid \theta_t] \leq (1 - \eta\mu(\gamma))(\widehat{\mathcal{L}}(\theta_t) - \widehat{\mathcal{L}}^*) + \frac{\beta(\gamma)\eta^2}{2} \sigma^2(\gamma).$$

Tomando expectativa total e iterando, obtém-se:

$$\mathbb{E}[\widehat{\mathcal{L}}(\theta_t) - \widehat{\mathcal{L}}^*] \leq (1 - \eta\mu(\gamma))^t \Delta_0 + \frac{\eta\sigma^2(\gamma)}{2\mu(\gamma)}.$$

**Passo 3 (substituir formas explícitas).** Use  $\eta = 1/\beta(\gamma)$ ,  $\mu(\gamma) \geq \alpha_L^2 \mu_L(0)$ ,  $\beta(\gamma) = C_\ell(\alpha_L^2 \beta_L + \alpha_H^2 \beta_H)$  e  $\sigma^2(\gamma) \leq C'_\ell(\alpha_L^2 \sigma_L^2 + \alpha_H^2 \sigma_H^2)$ . Dividindo numerador e denominador por  $\alpha_L^2$  aparece o termo  $\left(\frac{\alpha_H}{\alpha_L}\right)^2 = (1 - \gamma)^{2(\underline{w}_H - \underline{w}_L)}$  explicitamente.  $\square$

## 9.7 Discussão: quando o ruído acelera e quando ele destrói a bacia PL

**(i) Mecanismo matemático do “benefício” na convergência.** Os Teoremas 5–6 mostram que a razão

$$\left(\frac{\alpha_H(\gamma)}{\alpha_L(\gamma)}\right)^2 = (1 - \gamma)^{2(\underline{w}_H - \underline{w}_L)}$$

multiplica precisamente os termos de alta coerência associados a (i) maior suavidade adversa ( $\beta_H$ ) e (ii) maior variância ( $\sigma_H^2$ ). Como  $\underline{w}_H - \underline{w}_L > 0$ , aumentar  $\gamma$  reduz esses termos, melhorando:

- o número de condição efetivo  $\kappa(\gamma)$  (GD);
- o piso de ruído do SGD  $\frac{\eta\sigma^2(\gamma)}{2\mu(\gamma)}$  (SGD).

Isso formaliza a intuição: ruído “mata” componentes de alta coerência que tornam a paisagem irregular e ruidosa.

(ii) **Regime de falha (ruído excessivo).** O ganho requer que a região PL local  $\mathcal{B}$  continue válida e que  $\mu(\gamma) > 0$  não colapse. Se  $\gamma$  for grande, mesmo o bloco low sofre contração ( $\alpha_L(\gamma) \downarrow 0$ ), podendo:

- reduzir a informação do sinal e “achatar” demais a função;
- deslocar  $\theta_\gamma^*$  para fora de  $\mathcal{B}$ , quebrando a hipótese PL local.

Isso é a versão matemática do fato experimental: existe um *sweet spot*  $\gamma$ .

(iii) **Como usar na banca: constantes observáveis.** Os objetos a estimar/relatar experimentalmente:

$$\underline{w}_L, \underline{w}_H \Rightarrow (1 - \gamma)^{2(\underline{w}_H - \underline{w}_L)},$$

e proxies para  $\beta_L, \beta_H, \mu_L(0), \sigma_L^2, \sigma_H^2$  via: curvatura local (norma da Hessiana/estimadores), norma do gradiente e variância do gradiente estocástico. Isso transforma o “benefício” em uma afirmação verificável quantitativamente.

## 10 (3) Otimização de $\gamma$ e contraprova: por que regularizações isotrópicas não reproduzem o benefício

### 10.1 Conjunto admissível de ruído e variável reduzida

Os resultados da Seção (2) valem *condicionalmente* a duas restrições práticas: (i) a região PL local  $\mathcal{B}$  permanece válida (iterados não escapam), e (ii) o sinal low não colapsa (senão a tarefa perde margem informacional).

**Definição 6** (Conjunto admissível  $\Gamma_{\text{adm}}$ ). Fixe  $\gamma_{\text{max}} \in (0, 1)$  tal que, para todo  $\gamma \in [0, \gamma_{\text{max}}]$ , a PL local e a suavidade local valem em  $\mathcal{B}$ , e os iterados permanecem em  $\mathcal{B}$ . Opcionalmente, imponha um piso de amplitude no bloco low:

$$\alpha_L(\gamma)^2 = (1 - \gamma)^{2\underline{w}_L} \geq \underline{\alpha}^2 \quad (\underline{\alpha} \in (0, 1]).$$

Defina então

$$\Gamma_{\text{adm}} := \left\{ \gamma \in [0, \gamma_{\text{max}}] : (1 - \gamma)^{2\underline{w}_L} \geq \underline{\alpha}^2 \right\}.$$

Para explicitar a dependência em  $\gamma$  de forma limpa, introduzimos

$$u := 1 - \gamma \in (0, 1], \quad \Delta w := \underline{w}_H - \underline{w}_L > 0, \quad r(u) := \left( \frac{\alpha_H(\gamma)}{\alpha_L(\gamma)} \right)^2 = u^{2\Delta w} \in (0, 1].$$

Assim, tudo que na Seção (2) dependia de  $\gamma$  via  $\left( \frac{\alpha_H}{\alpha_L} \right)^2$  passa a depender de  $r(u) = u^{2\Delta w}$ .

## 10.2 (3.1) Corolário: $\gamma$ ótimo de convergência para GD (monotonicidade no regime admissível)

Defina as constantes (da Seção (2)):

$$\beta(\gamma) = C_\ell(\alpha_L^2\beta_L + \alpha_H^2\beta_H) = C_\ell\alpha_L^2(\beta_L + \beta_H r), \quad \mu(\gamma) \geq \alpha_L^2\mu_L(0).$$

A taxa linear do GD (com  $\eta = 1/\beta(\gamma)$ ) é controlada por

$$\rho(\gamma) := \frac{\mu(\gamma)}{\beta(\gamma)} \gtrsim \frac{\alpha_L^2\mu_L(0)}{C_\ell\alpha_L^2(\beta_L + \beta_H r)} = \frac{\mu_L(0)}{C_\ell(\beta_L + \beta_H r)}.$$

Observe que o fator  $\alpha_L^2$  cancela: *o ganho de convergência do GD é governado primariamente pela supressão relativa do bloco high via  $r = u^{2\Delta w}$ .*

**Corolário 1** (GD: escolha  $\gamma$  que maximiza  $\rho(\gamma)$  em  $\Gamma_{\text{adm}}$ ). *No regime admissível  $\Gamma_{\text{adm}}$ , a função*

$$\rho(\gamma) = \frac{\mu_L(0)}{C_\ell(\beta_L + \beta_H(1 - \gamma)^{2\Delta w})}$$

*é monótona crescente em  $\gamma$ . Logo, a melhor taxa de convergência linear (menor fator de contração por iteração) é atingida em*

$$\gamma_{\text{GD}}^* = \max \Gamma_{\text{adm}}.$$

*Proof. Passo 1 (reduzir ao parâmetro  $u = 1 - \gamma$ ). Como  $r(u) = u^{2\Delta w}$  e  $\Delta w > 0$ , temos  $\frac{dr}{du} = 2\Delta w u^{2\Delta w - 1} > 0$  para  $u \in (0, 1]$ .*

**Passo 2 (monotonicidade em  $r$ ).** Escreva

$$\rho(r) = \frac{\mu_L(0)}{C_\ell(\beta_L + \beta_H r)}.$$

Derivando em  $r$ :

$$\frac{d\rho}{dr} = -\frac{\mu_L(0)\beta_H}{C_\ell(\beta_L + \beta_H r)^2} < 0.$$

Logo  $\rho$  decresce quando  $r$  cresce.

**Passo 3 (compor monotonicidades).** Quando  $\gamma$  cresce,  $u = 1 - \gamma$  decresce, então  $r(u) = u^{2\Delta w}$  decresce. Como  $\rho$  decresce em  $r$ , conclui-se que  $\rho$  cresce em  $\gamma$ .

**Passo 4 (otimalidade no extremo admissível).** Sendo  $\rho$  crescente em  $\gamma$  em  $\Gamma_{\text{adm}}$ , o máximo ocorre em  $\max \Gamma_{\text{adm}}$ .  $\square$

Este corolário formaliza um ponto importante para banca: *no GD, dentro do intervalo onde a hipótese PL local não quebra, aumentar  $\gamma$  só ajuda a taxa*, pois reduz a curvatura adversa/high-frequency (capturada por  $\beta_H$ ) sem penalizar a razão  $\mu/\beta$  (o  $\alpha_L^2$  cancela). A barreira real é *geométrica/estatística* (saída de  $\mathcal{B}$ , colapso de margem, etc.), não algébrica.

### 10.3 (3.2) Teorema: $\gamma$ ótimo para SGD (trade-off entre supressão high e piso estocástico)

Para SGD, o termo de piso (Seção (2)) com passo  $\eta = 1/\beta(\gamma)$  é

$$\text{Floor}(\gamma) := \frac{\eta \sigma^2(\gamma)}{2\mu(\gamma)} \lesssim \frac{1}{2} \cdot \frac{\sigma^2(\gamma)}{\mu(\gamma)\beta(\gamma)}.$$

Usando as formas escalonadas:

$$\begin{aligned} \sigma^2(\gamma) &\leq C'_\ell(\alpha_L^2 \sigma_L^2 + \alpha_H^2 \sigma_H^2) = C'_\ell \alpha_L^2 (\sigma_L^2 + \sigma_H^2 r), \\ \mu(\gamma) &\geq \alpha_L^2 \mu_L(0), \quad \beta(\gamma) = C_\ell \alpha_L^2 (\beta_L + \beta_H r). \end{aligned}$$

Portanto,

$$\text{Floor}(\gamma) \leq \frac{C'_\ell}{2C_\ell \mu_L(0)} \cdot \frac{\sigma_L^2 + \sigma_H^2 r}{\alpha_L^2 (\beta_L + \beta_H r)}.$$

E como  $\alpha_L^2 = u^{2\underline{w}_L}$  e  $r = u^{2\Delta w}$ , definimos o funcional reduzido:

$$F(u) := \frac{\sigma_L^2 + \sigma_H^2 u^{2\Delta w}}{u^{2\underline{w}_L} (\beta_L + \beta_H u^{2\Delta w})}, \quad u \in (0, 1].$$

Minimizar o piso equivale a minimizar  $F(u)$  no conjunto admissível (em  $u$ ).

**Teorema 7** (SGD: candidato fechado para  $\gamma^*$  via equação quadrática em  $s = u^{2\Delta w}$ ). *Fixe  $\underline{w}_L \geq 1$  e  $\Delta w > 0$ , e defina*

$$k := \frac{\underline{w}_L}{\Delta w} > 0, \quad a := \sigma_L^2, \quad b := \sigma_H^2, \quad c := \beta_L, \quad d := \beta_H.$$

*Considere  $s := u^{2\Delta w} \in (0, 1]$  e reescreva*

$$F(u) \equiv \tilde{F}(s) = \frac{a + bs}{s^k(c + ds)}.$$

*Se existir  $s^* \in (0, 1]$  tal que  $\tilde{F}'(s^*) = 0$ , então  $s^*$  satisfaz a equação quadrática*

$$k b d s^2 + \left( d a (k + 1) + b c (k - 1) \right) s + k a c = 0.$$

*Qualquer raiz real positiva  $s^* \in (0, 1]$  fornece um candidato*

$$u^* = (s^*)^{\frac{1}{2\Delta w}}, \quad \gamma_{\text{SGD}}^* = 1 - u^*,$$

*e o minimizador global admissível é*

$$\gamma_{\text{SGD}}^* \in \arg \min_{\gamma \in \Gamma_{\text{adm}}} \text{Floor}(\gamma),$$

*isto é, ou  $\gamma_{\text{SGD}}^*$  coincide com esse candidato interior, ou ocorre em uma das bordas de  $\Gamma_{\text{adm}}$  (quando não há raiz admissível).*

*Proof. Passo 1 (reduzir em  $s$ ).* Com  $s = u^{2\Delta w}$ , temos  $u^{2\underline{w}L} = u^{2\Delta w \cdot k} = s^k$ . Logo

$$F(u) = \frac{a + bs}{s^k(c + ds)} = \tilde{F}(s), \quad s \in (0, 1].$$

**Passo 2 (derivar  $\log \tilde{F}$  para obter condição de estacionariedade).**

Derive

$$\log \tilde{F}(s) = \log(a + bs) - k \log s - \log(c + ds).$$

Então  $\tilde{F}'(s) = 0$  equivale a

$$\frac{b}{a + bs} - \frac{k}{s} - \frac{d}{c + ds} = 0.$$

**Passo 3 (simplificar o lado esquerdo: cancelamento chave).**

Traga os dois termos racionais para o mesmo denominador:

$$\frac{b}{a + bs} - \frac{d}{c + ds} = \frac{b(c + ds) - d(a + bs)}{(a + bs)(c + ds)} = \frac{bc - da}{(a + bs)(c + ds)}.$$

Logo a condição vira

$$\frac{bc - da}{(a + bs)(c + ds)} = \frac{k}{s}.$$

**Passo 4 (obter a quadrática).** Multiplicando por  $s(a + bs)(c + ds)$ :

$$(bc - da)s = k(a + bs)(c + ds) = k(ac + (ad + bc)s + bd s^2).$$

Reorganizando:

$$0 = kbd s^2 + (k(ad + bc) - (bc - da))s + kac.$$

E como

$$k(ad + bc) - (bc - da) = ad(k + 1) + bc(k - 1),$$

obtemos exatamente

$$k b d s^2 + \left( d a (k + 1) + b c (k - 1) \right) s + k a c = 0.$$

**Passo 5 (converter para  $\gamma$ ).** Dada raiz admissível  $s^* \in (0, 1]$ , defina  $u^* = (s^*)^{1/(2\Delta w)}$  e  $\gamma^* = 1 - u^*$ . Se não houver raiz admissível, o mínimo de  $\tilde{F}$  em um intervalo fechado admissível ocorre em borda.  $\square$

[Interpretação do trade-off do SGD] O piso do SGD carrega um fator *desfavorável*  $1/\alpha_L^2 = u^{-2\underline{w}L}$ : ruído grande demais (u pequeno) eleva o piso porque reduz a amplitude do gradiente do sinal e piora a relação sinal-ruído. Ao mesmo tempo, o termo  $r = u^{2\Delta w}$  reduz a parcela high (variância e curvatura adversa). O Teorema 7 formaliza exatamente este compromisso:

- Se a parte high domina (grande  $\sigma_H^2$  e/ou grande  $\beta_H$ ), tende a existir um  $\gamma^* > 0$  interior.
- Se a parte low domina (pequeno ganho ao suprimir high), o melhor pode ser  $\gamma$  próximo da borda de menor ruído.

#### 10.4 (3.3) Contraprova: regularização isotrópica (sem separação por $dw(P)$ ) não melhora $\kappa(\gamma)$ e pode piorar o piso do SGD

Agora mostramos uma limitação matemática: se o “ruído” ou regularização atua *isotropicamente* (mesmo fator para todos os modos), não há mecanismo para reduzir *relativamente* a parte high em relação à low. Logo o efeito de “suavização benéfica” desaparece.

**Definição 7** (Modelo alternativo isotrópico). Considere um operador de regularização/noise que induz

$$q_P(\gamma) \equiv q(\gamma) \in (0, 1] \quad \text{para todo } P \in \mathcal{P},$$

isto é, não depende de  $dw(P)$ . Assim,  $\alpha_L(\gamma) = \alpha_H(\gamma) = q(\gamma)$  e  $\left(\frac{\alpha_H}{\alpha_L}\right)^2 \equiv 1$ .

**Teorema 8** (Contraprova: isotropia implica nenhuma melhora de condicionamento e piso do SGD piora). *No modelo isotrópico, suponha que, localmente:*

$$\mu(\gamma) = q(\gamma)^2 \mu(0), \quad \beta(\gamma) = q(\gamma)^2 \beta(0), \quad \sigma^2(\gamma) = q(\gamma)^2 \sigma^2(0).$$

Então:

1. (GD) A razão efetiva  $\rho(\gamma) = \mu(\gamma)/\beta(\gamma)$  é invariante em  $\gamma$ :

$$\frac{\mu(\gamma)}{\beta(\gamma)} = \frac{\mu(0)}{\beta(0)}.$$

Logo, o ruído isotrópico não acelera o GD por condicionamento.

2. (SGD) Com passo  $\eta = 1/\beta(\gamma)$ , o piso em expectativa cresce como  $1/q(\gamma)^2$ :

$$\text{Floor}(\gamma) = \frac{\eta \sigma^2(\gamma)}{2\mu(\gamma)} = \frac{\sigma^2(0)}{2\mu(0)\beta(0)} \cdot \frac{1}{q(\gamma)^2},$$

portanto piora monotonicamente quando  $q(\gamma) \downarrow 0$ .

*Proof.* **Item 1 (GD).** Sob isotropia,  $\mu(\gamma) = q^2\mu(0)$  e  $\beta(\gamma) = q^2\beta(0)$ , logo

$$\frac{\mu(\gamma)}{\beta(\gamma)} = \frac{q^2\mu(0)}{q^2\beta(0)} = \frac{\mu(0)}{\beta(0)}.$$

Não há dependência em  $\gamma$ , então não há melhora de condicionamento induzida por ruído.

**Item 2 (SGD).** Com  $\eta = 1/\beta(\gamma) = 1/(q^2\beta(0))$  e  $\sigma^2(\gamma) = q^2\sigma^2(0)$ :

$$\text{Floor}(\gamma) = \frac{\eta\sigma^2(\gamma)}{2\mu(\gamma)} = \frac{1}{q^2\beta(0)} \cdot \frac{q^2\sigma^2(0)}{2q^2\mu(0)} = \frac{\sigma^2(0)}{2\mu(0)\beta(0)} \cdot \frac{1}{q^2}.$$

Como  $q(\gamma) \leq 1$ , o piso não melhora e tipicamente piora com ruído.  $\square$

[Por que isto refuta o modelo alternativo] A descoberta de “ruído benéfico” depende criticamente de um *filtro espectral* por  $dw(P)$ :

$$q_P(\gamma) = (1 - \gamma)^{dw(P)} \Rightarrow \frac{\alpha_H}{\alpha_L} = (1 - \gamma)^{\Delta w} < 1,$$

ou seja, o high é suprimido *mais* do que o low. No alternativo isotrópico,  $\alpha_H = \alpha_L$ , então inexistente esse canal seletivo. A consequência matemática é objetiva:

- GD: sem melhora de  $\mu/\beta$ , não há aceleração por “alisamento”.
- SGD: a amplitude global cai, mas a variância relativa não cai na proporção necessária, elevando o piso.

Portanto, um modelo alternativo que não codifica a dependência em  $dw(P)$  não reproduz o fenômeno.

## 11 (4) Verificabilidade, generalização e robustez: como “fechar” a validação matemática em banca

### 11.1 Objetivo da seção e conexão com as seções (1)–(3)

As Seções (1)–(3) estabeleceram:

- (A) **existência de  $\gamma^* > 0$  benéfico** sob (i) sobreparametrização, (ii) amostra finita e (iii) *componente espúria em coerências/high-frequency*;
- (2) **convergência (GD/SGD)** sob **PL local** no *landscape suavizado pelo ruído seletivo*  $q_P(\gamma) = (1 - \gamma)^{dw(P)}$ ;

- (3) **otimização de  $\gamma$  e contraprova** refutando regularização isotrópica.

Falta agora tornar o modelo *auditável*: (i) como estimar/justificar quantitativamente as constantes do modelo, (ii) como provar uma propriedade de generalização/robustez coerente com a interpretação de “regularização benéfica”, e (iii) como apresentar um protocolo metodológico que uma banca possa reproduzir e criticar.

## 11.2 (4.1) Definições: risco, empiria, estabilidade e robustez

Considere um conjunto de treino  $S = \{z_i = (x_i, y_i)\}_{i=1}^m$  i.i.d. de uma distribuição  $\mathcal{D}$ . Seja  $f_\theta$  o classificador quântico variacional (saída real escalar) e  $\ell(\cdot, \cdot)$  uma perda de classificação (por exemplo, logistic, hinge suavizada, cross-entropy).

**Definição 8** (Risco verdadeiro e risco empírico).

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

**Definição 9** (Algoritmo e saída). Denote por  $\mathcal{A}$  o algoritmo de treino (GD/SGD) que, dado  $S$  e um ruído  $\gamma$ , devolve

$$\theta_T = \mathcal{A}(S; \gamma, T).$$

**Definição 10** (Estabilidade uniforme (no parâmetro)). Seja  $S^{(j)}$  o dataset em que o exemplo  $z_j$  foi substituído por  $z'_j$  i.i.d. de  $\mathcal{D}$ . Dizemos que  $\mathcal{A}$  é  $\varepsilon_{\text{stab}}$ -estável (no parâmetro) se

$$\mathbb{E} [\|\mathcal{A}(S; \gamma, T) - \mathcal{A}(S^{(j)}; \gamma, T)\|_2] \leq \varepsilon_{\text{stab}} \quad \text{para todo } j.$$

**Definição 11** (Robustez (na predição)). Assuma que  $f_\theta$  é  $L_f$ -Lipschitz em  $\theta$  (localmente na bola  $\mathcal{B}$ ):

$$|f_\theta(x) - f_{\theta'}(x)| \leq L_f \|\theta - \theta'\|_2.$$

Então a robustez de predição induzida pela estabilidade é

$$\mathbb{E} [|f_{\theta_T}(x) - f_{\theta'_T}(x)|] \leq L_f \varepsilon_{\text{stab}}.$$

*Discussão/Observação 11.* A estabilidade (no sentido de troca de um exemplo) é o caminho metodologicamente mais aceito em banca para conectar: *ruído*  $\rightarrow$  *suavização do treino*  $\rightarrow$  *menor sensibilidade a dados*  $\rightarrow$  *melhor generalização*.

### 11.3 (4.2) Lema: Lipschitz efetivo sob ruído seletivo (dependência em $(1 - \gamma)^{dw(P)}$ )

Recall da decomposição low/high da Seção (2): existe  $\underline{w}_L < \underline{w}_H$  tal que, para todo  $P \in \mathcal{P}_L$ ,  $dw(P) \leq \underline{w}_L$ , e para todo  $P \in \mathcal{P}_H$ ,  $dw(P) \geq \underline{w}_H$ . Defina

$$\alpha_L(\gamma) = (1 - \gamma)^{\underline{w}_L}, \quad \alpha_H(\gamma) = (1 - \gamma)^{\underline{w}_H}.$$

**Lema 7** (Contração seletiva do Jacobiano em função de  $\gamma$ ). *Assuma que, localmente em  $\mathcal{B}$ , o mapa  $\theta \mapsto f_\theta(x)$  admite decomposição linear em operadores de medição com coeficientes dependentes de  $\theta$  de forma suave, e que o ruído seletivo atua multiplicativamente em cada modo  $P$  por  $q_P(\gamma)$ . Então existe  $C_J > 0$  e uma decomposição do Jacobiano  $\nabla_\theta f_\theta(x)$  em componentes low/high tal que*

$$\|\nabla_\theta f_\theta(x)\|_2 \leq C_J \left( \alpha_L(\gamma) \|\nabla_\theta f_\theta(x)\|_L + \alpha_H(\gamma) \|\nabla_\theta f_\theta(x)\|_H \right).$$

*Em particular, se a parte espúria domina o Jacobiano (isto é,  $\|\nabla f\|_H \gg \|\nabla f\|_L$ ), a redução relativa  $\alpha_H(\gamma)/\alpha_L(\gamma) = (1 - \gamma)^{\Delta w}$  diminui o “Lipschitz efetivo” da predição.*

*Proof.* **Passo 1 (linearidade do efeito do canal em modos).** Por hipótese, cada contribuição associada a um modo  $P$  é multiplicada por  $q_P(\gamma)$ . Como  $q_P(\gamma) \leq \alpha_L(\gamma)$  em  $\mathcal{P}_L$  e  $q_P(\gamma) \leq \alpha_H(\gamma)$  em  $\mathcal{P}_H$ , a soma (ou norma) separa-se em duas parcelas majoradas.

**Passo 2 (agregação via desigualdade triangular + constante estrutural).** Agrupe termos low/high. A passagem de norma do vetor agregado para soma de normas de blocos introduz um fator geométrico  $C_J$  dependente apenas da base/parametrização local. Conclui-se a desigualdade.  $\square$

Este lema é o elo formal que faltava: ele mostra como o ruído seletivo reduz a sensibilidade do modelo a perturbações em  $\theta$  *principalmente* onde havia energia espúria (high). Isso prepara a prova de estabilidade e, portanto, de generalização.

### 11.4 (4.3) Teorema: generalização via estabilidade sob PL local (com dependência explícita em $(1 - \gamma)^{dw(P)}$ )

Vamos assumir um resultado padrão de estabilidade para SGD/GD em funções suaves com uma forma de convexidade efetiva (aqui: PL local). Para manter a banca *matematicamente satisfeita*, enunciamos uma versão auto-contida como hipótese de trabalho.

**Hipótese 10** (PL local + suavidade + gradiente limitado no tubo). Existe uma bola  $\mathcal{B}$  tal que, para todo  $\theta \in \mathcal{B}$ : (i)  $\widehat{R}_S(\theta)$  satisfaz PL local com constante  $\mu(\gamma)$ , (ii)  $\widehat{R}_S$  é  $\beta(\gamma)$ -suave, (iii)  $\|\nabla \ell(f_\theta(x), y)\|_2 \leq G(\gamma)$  quase certamente, e (iv) o algoritmo permanece em  $\mathcal{B}$  para  $t \leq T$ .

**Teorema 9** (Gap de generalização controlado por estabilidade, com ganho seletivo em  $\gamma$ ). *Sob a Hipótese 10 e com passo constante  $\eta \leq 1/\beta(\gamma)$ , considere SGD com amostragem uniforme de exemplos e ruído seletivo  $q_P(\gamma)$ . Então existe constante universal  $C > 0$  (independente de  $m$ ) tal que o gap esperado satisfaz*

$$\mathbb{E}[R(\theta_T) - \widehat{R}_S(\theta_T)] \leq C \cdot \frac{\eta T}{m} G(\gamma)^2.$$

Além disso, o termo  $G(\gamma)$  admite a decomposição (via Lema 7 e Lipschitz de  $\ell$ ):

$$G(\gamma) \leq C_\ell'' \left( \alpha_L(\gamma) G_L + \alpha_H(\gamma) G_H \right),$$

com  $G_H \gg G_L$  quando há componente espúria dominante. Logo, no regime onde  $G_H$  domina, o bound melhora com  $\gamma$  aproximadamente como

$$\mathbb{E}[R(\theta_T) - \widehat{R}_S(\theta_T)] \lesssim \frac{\eta T}{m} \alpha_H(\gamma)^2 G_H^2 = \frac{\eta T}{m} (1 - \gamma)^{2w_H} G_H^2,$$

até que o termo low (proporcional a  $\alpha_L(\gamma) G_L$ ) torne-se dominante.

*Proof. Passo 1 (recorrência de estabilidade em SGD).* Considere duas execuções acopladas do SGD, uma em  $S$  e outra em  $S^{(j)}$ , com o mesmo stream de mini-batches, diferindo apenas quando o exemplo  $j$  é selecionado. O passo SGD é

$$\theta_{t+1} = \theta_t - \eta g(\theta_t; z_{i_t}), \quad \theta'_{t+1} = \theta'_t - \eta g(\theta'_t; z'_{i_t}),$$

onde  $g(\theta; z) = \nabla_\theta \ell(f_\theta(x), y)$ .

**Passo 2 (majorar a diferença por suavidade).** Pela suavidade,  $g(\cdot; z)$  é Lipschitz com constante  $\beta(\gamma)$  (localmente):

$$\|g(\theta; z) - g(\theta'; z)\| \leq \beta(\gamma) \|\theta - \theta'\|.$$

Quando  $z_{i_t}$  coincide em ambas as execuções, obtemos:

$$\|\theta_{t+1} - \theta'_{t+1}\| \leq (1 + \eta\beta(\gamma)) \|\theta_t - \theta'_t\|.$$

Quando o exemplo difere (probabilidade  $1/m$  por iteração sob amostragem uniforme), usamos o gradiente limitado:

$$\|g(\theta; z) - g(\theta'; z')\| \leq \|g(\theta; z)\| + \|g(\theta'; z')\| \leq 2G(\gamma),$$

logo

$$\|\theta_{t+1} - \theta'_{t+1}\| \leq \|\theta_t - \theta'_t\| + 2\eta G(\gamma).$$

**Passo 3 (tomar expectativa e resolver a recorrência).** Tomando expectativa (no stream e na troca  $S \rightarrow S^{(j)}$ ), aparece um termo  $\frac{1}{m}$  multiplicando o “choque”  $2\eta G(\gamma)$ . A solução padrão da recorrência fornece

$$\mathbb{E}\|\theta_T - \theta'_T\| \leq C \cdot \frac{\eta T}{m} G(\gamma),$$

para alguma constante  $C$  quando  $\eta \leq 1/\beta(\gamma)$  e dentro do tubo local.

**Passo 4 (converter estabilidade em gap).** Se  $\ell$  é Lipschitz em  $\theta$  com constante proporcional a  $G(\gamma)$ , então

$$\mathbb{E}[R(\theta_T) - \hat{R}_S(\theta_T)] \leq \varepsilon_{\text{stab}} \cdot G(\gamma) \lesssim \frac{\eta T}{m} G(\gamma)^2,$$

que prova a primeira desigualdade.

**Passo 5 (decomposição seletiva de  $G(\gamma)$ ).** A partir do Lema 7 e Lipschitz de  $\ell$  na saída  $f_\theta(x)$ , obtemos o bound de  $G(\gamma)$  em termos de  $\alpha_L, \alpha_H$ . Inserindo, resulta a forma final.  $\square$

[Leitura prática para banca] O Teorema 9 torna *matematicamente defensável* o argumento “ruído ajuda generalização”: o ruído seletivo reduz o tamanho do gradiente espúrio  $G_H$  por um fator  $(1 - \gamma)^{\underline{w}_H}$ , o que reduz o bound do gap de generalização. O ponto  $\gamma^*$  surge quando:

- aumentar  $\gamma$  ainda reduz  $G_H$  (bom),
- mas começa a reduzir demais o componente low  $G_L$  e/ou viola a admissibilidade (PL local, margem, permanência em  $\mathcal{B}$ ).

Isso coincide com o comportamento “sweet spot” observado empiricamente.

## 11.5 (4.4) Proposição: como estimar as constantes do modelo a partir do experimento

A banca vai exigir que  $\underline{w}_L, \underline{w}_H, \beta_L, \beta_H, \sigma_L^2, \sigma_H^2, \mu(\gamma)$  sejam (i) definidas, (ii) mensuráveis e (iii) reportáveis.

**Definição 12** (Partição operacional low/high por peso de Pauli). Fixe um limiar  $w_0$ . Defina:

$$\mathcal{P}_L = \{P : dw(P) \leq w_0\}, \quad \mathcal{P}_H = \{P : dw(P) > w_0\}.$$

Tome  $\underline{w}_L := w_0$  e  $\underline{w}_H := w_0 + 1$  (ou uma margem maior, se desejado).

**Proposição 2** (Estimadores consistentes de energia de gradiente e variância por blocos). *Considere SGD com mini-batch de tamanho  $b$  e gradiente estocástico  $g_t$ . Suponha que existe um operador de projeção  $\Pi_L, \Pi_H$  no espaço de modos (implementado por decomposição em Pauli-strings/observáveis). Defina estimadores empíricos:*

$$\begin{aligned} \hat{G}_L^2 &:= \frac{1}{T} \sum_{t=1}^T \|\Pi_L g_t\|^2, & \hat{G}_H^2 &:= \frac{1}{T} \sum_{t=1}^T \|\Pi_H g_t\|^2, \\ \hat{\sigma}_L^2 &:= \frac{1}{T} \sum_{t=1}^T \|\Pi_L(g_t - \bar{g})\|^2, & \hat{\sigma}_H^2 &:= \frac{1}{T} \sum_{t=1}^T \|\Pi_H(g_t - \bar{g})\|^2, \end{aligned}$$

onde  $\bar{g} = \frac{1}{T} \sum_{t=1}^T g_t$ . Se  $\{g_t\}$  for ergódica/estacionária no tubo  $\mathcal{B}$  (após burn-in), então, quando  $T \rightarrow \infty$ ,

$$\hat{G}_L^2 \rightarrow \mathbb{E}\|\Pi_L g\|^2, \quad \hat{G}_H^2 \rightarrow \mathbb{E}\|\Pi_H g\|^2, \quad \hat{\sigma}_L^2 \rightarrow \text{Var}(\Pi_L g), \quad \hat{\sigma}_H^2 \rightarrow \text{Var}(\Pi_H g),$$

em probabilidade.

**Proof. Passo 1 (lei dos grandes números para sequências estacionárias).**

Sob ergodicidade/estacionariedade, médias temporais convergem para médias em distribuição.

**Passo 2 (projeção é linear e limitada).** Como  $\Pi_L, \Pi_H$  são lineares e de norma  $\leq 1$ , preservam integrabilidade necessária.

**Passo 3 (aplicar LLN/ergodic theorem).** Aplica-se o teorema ergódico a  $\|\Pi_L g_t\|^2$  e  $\|\Pi_H(g_t - \bar{g})\|^2$ .  $\square$

[Como operacionalizar  $\Pi_L, \Pi_H$  no seu caso] Em Qiskit/PennyLane, a decomposição em Pauli-strings aparece de forma natural:

- no *modelo de ruído*, o canal atua em operadores (Heisenberg picture) e permite rastrear a atenuação de cada Pauli-string;
- o  $dw(P)$  pode ser aproximado por *peso de Pauli* (número de qubits com  $X/Y/Z$ ) ou por *light-cone* (propagação sob conjugação);

- o limiar  $w_0$  pode ser fixado para separar termos com maior probabilidade de serem espúrios (altos pesos) em arquitetura sobreparametrizada.

A banca não exige que  $\Pi_L$  seja perfeito; exige que seja *reprodutível*, *justificado* e *sensível* ao efeito do ruído.

## 11.6 (4.5) Checklist metodológico para banca (procedimento reproduzível)

[Roteiro de validação matemática e experimental] **Entrada:** arquitetura VQC, dataset, perda  $\ell$ , algoritmo (GD/SGD), família de ruído seletivo  $q_P(\gamma) = (1 - \gamma)^{dw(P)}$ .

**Saída:** teoremas instanciados + estimativas de constantes + contraprova.

1. **Especificar a decomposição low/high:** escolher  $w_0$ , definir  $\mathcal{P}_L, \mathcal{P}_H$ , reportar  $\underline{w}_L, \underline{w}_H$  e justificar via arquitetura (depth, conectividade, pesos de Pauli).
2. **Verificar admissibilidade  $\Gamma_{\text{adm}}$ :** rodar experimento piloto e checar (i) permanência em  $\mathcal{B}$  (norma do gradiente e parâmetro), (ii) ausência de colapso de margem (acurácia não se aproxima de aleatório), (iii) regimes em que PL local é plausível (curva de perda e não explosão de Hessiana).
3. **Estimar blocos:** usar a Proposição 2 para obter  $\hat{G}_L, \hat{G}_H, \hat{\sigma}_L^2, \hat{\sigma}_H^2$ .
4. **Estimar suavidade  $\beta(\gamma)$ :** aproximar  $\beta$  por norma espectral da Hessiana via diferenças finitas:

$$\hat{\beta}(\gamma) \approx \max_{r=1, \dots, R} \frac{\|\nabla \hat{R}(\theta + \delta v_r) - \nabla \hat{R}(\theta)\|}{\delta},$$

com vetores aleatórios  $v_r$  (normalizados) e  $\delta$  pequeno.

5. **Estimar PL local  $\mu(\gamma)$ :** no tubo, usar o quociente PL empírico:

$$\hat{\mu}(\gamma) := \min_{t \in \mathcal{T}} \frac{\|\nabla \hat{R}(\theta_t)\|^2}{2(\hat{R}(\theta_t) - \hat{R}_{\min})},$$

onde  $\hat{R}_{\min}$  é o menor valor observado (lower envelope).

6. **Instanciar os teoremas:** substituir constantes estimadas nas taxas de (2) e nos bounds de generalização de (4.3).

7. **Otimizar  $\gamma$** : usar (3.1)–(3.2) (GD/SGD) para prever  $\gamma^*$  e comparar com varredura ou BO.
8. **Contraprova**: executar o modelo isotrópico (Teorema 8) com mesma intensidade efetiva (mesmo  $q(\gamma)$  médio), e reportar que não há melhora de condicionamento e/ou o piso piora, como previsto.

O Procedimento 11.6 atende o padrão “matemática + metodologia”: todas as constantes do enunciado são (i) definidas, (ii) estimáveis por protocolo, (iii) conectadas ao comportamento observado e (iv) testadas contra um modelo alternativo refutado.

## 12 (5) Robustez adversarial / invariância induzida por ruído seletivo e Teorema-síntese

### 12.1 Motivação

Além de (A) *benefício condicionado* e (2)–(4) *convergência + generalização*, uma banca normalmente exige uma propriedade adicional do tipo:

“o mecanismo proposto não apenas otimiza melhor, mas *reduz sensibilidade* a perturbações”.

Nesta seção, formalizamos um **Teorema de robustez adversarial/invariância** decorrente do mesmo fator seletivo

$$q_P(\gamma) = (1 - \gamma)^{dw(P)}.$$

Ao final, consolidamos tudo em um **Teorema-síntese** (pronto para conclusão de artigo matemático).

### 12.2 (5.1) Definições: modelo quântico, embedding e noção de adversário

**Definição 13** (Embedding quântico e estado de entrada). Seja  $x \in \mathbb{R}^p$  uma amostra clássica. Um embedding quântico é um mapa

$$\rho_{\text{in}}(x) \in \mathcal{D}(\mathcal{H}_{2^n}),$$

onde  $\mathcal{D}$  é o conjunto de estados densidade em  $n$  qubits.

**Definição 14** (Classificador VQC sob ruído seletivo). Fixe um circuito parametrizado  $U_\theta$  e um observável (medida)  $M$  com  $M = M^\dagger$ . Defina o score sob ruído seletivo como

$$f_\theta^\gamma(x) := \text{Tr} \left[ M \mathcal{N}_\gamma(U_\theta \rho_{\text{in}}(x) U_\theta^\dagger) \right],$$

onde  $\mathcal{N}_\gamma$  é o canal de ruído seletivo (CPTP) que, na decomposição em Pauli-strings, atua multiplicando cada modo  $P$  por  $q_P(\gamma) = (1 - \gamma)^{dw(P)}$ .

**Definição 15** (Métrica de perturbação e adversário). Assuma que o adversário escolhe  $x'$  tal que  $\|x - x'\|_2 \leq \varepsilon$ . Dizemos que o modelo é  $L$ -robusto se

$$|f_\theta^\gamma(x) - f_\theta^\gamma(x')| \leq L\varepsilon \quad \text{para todo } x, x'.$$

**Hipótese 11** (Embedding Lipschitz em norma de Hilbert–Schmidt). Existe  $L_E > 0$  tal que, para todo  $x, x'$ ,

$$\|\rho_{\text{in}}(x) - \rho_{\text{in}}(x')\|_2 \leq L_E \|x - x'\|_2,$$

onde  $\|A\|_2 = \sqrt{\text{Tr}(A^\dagger A)}$ .

*Discussão/Observação 12.* A escolha de  $\|\cdot\|_2$  (em vez de traço  $\|\cdot\|_1$ ) não é “cosmética”: o canal seletivo contrai *coeficientes de Pauli* diretamente, e essa contração se traduz de forma limpa em norma 2.

### 12.3 (5.2) Lema: reescrita Heisenberg + bound por Cauchy–Schwarz

**Lema 8** (Forma Heisenberg e bound por norma 2). Defina o observável efetivo (Heisenberg) como

$$\widetilde{M}_{\theta,\gamma} := U_\theta^\dagger \mathcal{N}_\gamma^\dagger(M) U_\theta,$$

onde  $\mathcal{N}_\gamma^\dagger$  é o adjunto de Heisenberg do canal. Então, para quaisquer  $x, x'$ ,

$$f_\theta^\gamma(x) - f_\theta^\gamma(x') = \text{Tr} \left[ \widetilde{M}_{\theta,\gamma} (\rho_{\text{in}}(x) - \rho_{\text{in}}(x')) \right],$$

e vale o bound

$$|f_\theta^\gamma(x) - f_\theta^\gamma(x')| \leq \|\widetilde{M}_{\theta,\gamma}\|_2 \cdot \|\rho_{\text{in}}(x) - \rho_{\text{in}}(x')\|_2.$$

*Proof.* **Passo 1 (cíclica do traço + adjunto).** Pela definição do adjunto,

$$\mathrm{Tr}[M \mathcal{N}_\gamma(\sigma)] = \mathrm{Tr}[\mathcal{N}_\gamma^\dagger(M) \sigma].$$

Aplicando com  $\sigma = U_\theta \rho U_\theta^\dagger$  e usando ciclicidade do traço, obtém-se a forma com  $\widetilde{M}_{\theta, \gamma}$ .

**Passo 2 (Cauchy–Schwarz em Hilbert–Schmidt).** Usa-se  $|\mathrm{Tr}(A^\dagger B)| \leq \|A\|_2 \|B\|_2$  com  $A = \widetilde{M}_{\theta, \gamma}$  e  $B = \rho_{\mathrm{in}}(x) - \rho_{\mathrm{in}}(x')$ .  $\square$

O problema reduz-se a entender como  $\|\widetilde{M}_{\theta, \gamma}\|_2$  depende de  $\gamma$ . A seguir, exploramos a decomposição em Pauli-strings e a atenuação seletiva  $q_P(\gamma)$ .

## 12.4 (5.3) Lema: contração seletiva da norma 2 do observável efetivo

Considere a decomposição em base de Pauli:

$$A = \sum_{P \in \mathcal{P}} a_P P, \quad \text{com} \quad \langle P, Q \rangle := \mathrm{Tr}(P^\dagger Q) = 2^n \mathbf{1}\{P = Q\}.$$

Logo,

$$\|A\|_2^2 = \mathrm{Tr}(A^\dagger A) = 2^n \sum_P |a_P|^2.$$

**Lema 9** (Atenuação por  $q_P(\gamma)$  em norma 2). *Se  $\mathcal{N}_\gamma^\dagger$  atua em cada Pauli-string como*

$$\mathcal{N}_\gamma^\dagger(P) = q_P(\gamma) P, \quad q_P(\gamma) = (1 - \gamma)^{dw(P)},$$

*então para qualquer observável  $A = \sum_P a_P P$ ,*

$$\|\mathcal{N}_\gamma^\dagger(A)\|_2^2 = 2^n \sum_P |a_P|^2 q_P(\gamma)^2.$$

*Em particular, se definimos uma partição low/high por um limiar  $w_0$ ,*

$$\mathcal{P}_L = \{P : dw(P) \leq w_0\}, \quad \mathcal{P}_H = \{P : dw(P) > w_0\},$$

*e  $\underline{w}_L := w_0$ ,  $\underline{w}_H := w_0 + 1$ , então*

$$\|\mathcal{N}_\gamma^\dagger(A)\|_2 \leq \alpha_L(\gamma) \|A_L\|_2 + \alpha_H(\gamma) \|A_H\|_2,$$

*onde  $A_L = \sum_{P \in \mathcal{P}_L} a_P P$ ,  $A_H = \sum_{P \in \mathcal{P}_H} a_P P$  e*

$$\alpha_L(\gamma) = (1 - \gamma)^{\underline{w}_L}, \quad \alpha_H(\gamma) = (1 - \gamma)^{\underline{w}_H}.$$

*Proof. Passo 1 (ortogonalidade de Pauli-strings).* Como  $\text{Tr}(P^\dagger Q) = 0$  para  $P \neq Q$ , temos

$$\|\mathcal{N}_\gamma^\dagger(A)\|_2^2 = 2^n \sum_P |a_P q_P(\gamma)|^2.$$

**Passo 2 (separação low/high).** Em  $\mathcal{P}_L$ ,  $q_P(\gamma) \leq \alpha_L(\gamma)$ ; em  $\mathcal{P}_H$ ,  $q_P(\gamma) \leq \alpha_H(\gamma)$ . Logo,

$$\|\mathcal{N}_\gamma^\dagger(A)\|_2^2 \leq 2^n \left( \alpha_L(\gamma)^2 \sum_{P \in \mathcal{P}_L} |a_P|^2 + \alpha_H(\gamma)^2 \sum_{P \in \mathcal{P}_H} |a_P|^2 \right).$$

Tomando raiz e usando  $\sqrt{u+v} \leq \sqrt{u} + \sqrt{v}$ , obtém-se o bound linear em  $\|A_L\|_2, \|A_H\|_2$ .  $\square$

Este lema é o análogo, no espaço de observáveis, do “efeito de suavização” discutido para o gradiente. Ele formaliza: se o observável efetivo carrega energia em termos de alto peso (high), então a norma 2 contrai rapidamente com  $\gamma$ .

## 12.5 (5.4) Teorema: robustez adversarial/margem com constante explícita em $(1 - \gamma)^{dw(P)}$

**Teorema 10** (Robustez adversarial por contração seletiva). *Suponha a Hipótese 11. Considere o classificador  $f_\theta^\gamma$  e o observável efetivo  $\widetilde{M}_{\theta,\gamma} = U_\theta^\dagger \mathcal{N}_\gamma^\dagger(M) U_\theta$ . Então  $f_\theta^\gamma$  é  $L_{\text{adv}}(\gamma)$ -robusto com*

$$L_{\text{adv}}(\gamma) := L_E \|\widetilde{M}_{\theta,\gamma}\|_2.$$

Além disso, decompondo  $M_\theta := U_\theta^\dagger M U_\theta$  em low/high pela base de Pauli, vale o bound explícito:

$$L_{\text{adv}}(\gamma) \leq L_E \left( \alpha_L(\gamma) \|(M_\theta)_L\|_2 + \alpha_H(\gamma) \|(M_\theta)_H\|_2 \right),$$

com  $\alpha_L(\gamma) = (1 - \gamma)^{w_L}$  e  $\alpha_H(\gamma) = (1 - \gamma)^{w_H}$ . Em particular, se a parte espúria domina  $(M_\theta)_H$  (isto é,  $\|(M_\theta)_H\|_2 \gg \|(M_\theta)_L\|_2$ ), então

$$L_{\text{adv}}(\gamma) \approx L_E \alpha_H(\gamma) \|(M_\theta)_H\|_2 = L_E (1 - \gamma)^{w_H} \|(M_\theta)_H\|_2,$$

logo aumentar  $\gamma$  melhora a robustez (reduz  $L_{\text{adv}}$ ) até o ponto em que o termo low passa a dominar.

*Proof.* **Passo 1 (reduzir a diferença ao estado de entrada).** Pelo Lema 8,

$$|f_\theta^\gamma(x) - f_\theta^\gamma(x')| \leq \|\widetilde{M}_{\theta,\gamma}\|_2 \cdot \|\rho_{\text{in}}(x) - \rho_{\text{in}}(x')\|_2.$$

**Passo 2 (usar Lipschitz do embedding).** Pela Hipótese 11,

$$\|\rho_{\text{in}}(x) - \rho_{\text{in}}(x')\|_2 \leq L_E \|x - x'\|_2 \leq L_E \varepsilon.$$

Logo,

$$|f_\theta^\gamma(x) - f_\theta^\gamma(x')| \leq L_E \|\widetilde{M}_{\theta,\gamma}\|_2 \varepsilon,$$

o que define  $L_{\text{adv}}(\gamma)$ .

**Passo 3 (dependência em  $\gamma$  via decomposição em Pauli).** Como conjugação por  $U_\theta$  preserva a norma 2,  $\|U_\theta^\dagger A U_\theta\|_2 = \|A\|_2$ . Assim,  $\|\widetilde{M}_{\theta,\gamma}\|_2 = \|\mathcal{N}_\gamma^\dagger(M_\theta)\|_2$ . Aplicando o Lema 9 com  $A = M_\theta$ , obtém-se o bound low/high.  $\square$

[Interpretação por margem] Se o classificador final é  $\text{sign}(f_\theta^\gamma(x))$  e há **margem**  $\kappa > 0$  em um ponto  $x$ :

$$y f_\theta^\gamma(x) \geq \kappa,$$

então toda perturbação adversarial com  $\|x - x'\|_2 \leq \kappa/L_{\text{adv}}(\gamma)$  preserva o rótulo. Portanto, reduzir  $L_{\text{adv}}(\gamma)$  (via contração seletiva) aumenta o raio robusto. Isto complementa o resultado de (4.3): o mesmo mecanismo que reduz instabilidade e gap tende a aumentar invariância local.

## 12.6 (5.5) Corolário: trade-off robustez–acurácia e ponto $\gamma^*$

**Corolário 2** (Ponto ótimo  $\gamma^*$  como equilíbrio low/high). *No regime em que  $(M_\theta)_H$  domina para  $\gamma \approx 0$ , a robustez melhora monotonicamente em  $\gamma$  até que*

$$\alpha_H(\gamma) \|(M_\theta)_H\|_2 \approx \alpha_L(\gamma) \|(M_\theta)_L\|_2.$$

Defina  $\gamma^\dagger$  como solução aproximada de

$$(1 - \gamma)^{w_H} \|(M_\theta)_H\|_2 = (1 - \gamma)^{w_L} \|(M_\theta)_L\|_2.$$

Então

$$\gamma^\dagger \approx 1 - \left( \frac{\|(M_\theta)_L\|_2}{\|(M_\theta)_H\|_2} \right)^{\frac{1}{w_H - w_L}}.$$

Esse mesmo ponto é consistente com o  $\gamma^*$  das Seções (A) e (3): antes de  $\gamma^\dagger$  o ruído remove espúrio (benéfico), depois de  $\gamma^\dagger$  ele começa a degradar sinal low (prejudicial).

*Proof.* Rearranje a igualdade:

$$(1 - \gamma)^{\underline{w}_H - \underline{w}_L} = \frac{\|(M_\theta)_L\|_2}{\|(M_\theta)_H\|_2}.$$

Como  $\underline{w}_H - \underline{w}_L > 0$ , tomando potência  $1/(\underline{w}_H - \underline{w}_L)$  e isolando  $\gamma$  resulta a expressão.  $\square$

O corolário fornece um **estimador fechado** para o ponto de equilíbrio robustez-accurácia, análogo ao estimador de  $\gamma^*$  obtido por risco em (3). Isso é o tipo de “cálculo interpretável” que banca costuma aceitar muito bem.

## 12.7 (5.6) Teorema-síntese: benefício condicionado + convergência + generalização + robustez + contraprova

**Teorema 11** (Teorema-síntese do ruído quântico benéfico em VQCs). *Considere um classificador quântico variacional  $f_\theta^\gamma$  treinado em amostra finita  $S$ , sob ruído seletivo  $q_P(\gamma) = (1 - \gamma)^{dw(P)}$ .*

*Assuma:*

1. **(Overparameterization + espúrio em coerências)** existe decomposição low/high de modos (Pauli/light-cone) tal que a parcela high é majoritariamente espúria e domina a variância/sensibilidade para  $\gamma \approx 0$ ;
2. **(Admissibilidade)** existe um intervalo  $\Gamma_{\text{adm}} \subset (0, 1)$  onde (i) o algoritmo permanece no tubo local  $\mathcal{B}$ , (ii) o sinal low não colapsa, e (iii) as constantes locais de suavidade/PL são finitas;
3. **(PL local no landscape suavizado)** no tubo  $\mathcal{B}$  a perda empírica suavizada satisfaz PL local com constante  $\mu(\gamma) > 0$  e é  $\beta(\gamma)$ -suave, com  $\eta \leq 1/\beta(\gamma)$ ;
4. **(Embedding Lipschitz)** vale a Hipótese 11.

Então existe  $\gamma^* \in \Gamma_{\text{adm}}$  tal que, ao comparar  $\gamma = \gamma^*$  com  $\gamma = 0$ :

1. **(Benefício condicionado)** o risco (ou um proxy teórico do risco) diminui devido à supressão seletiva da parcela high espúria, com trade-off controlado pela atenuação do sinal low;
2. **(Convergência)** GD/SGD converge linearmente (em expectativa) no tubo:

$$\mathbb{E}[\widehat{R}(\theta_T) - \widehat{R}(\theta^*)] \leq (1 - \eta\mu(\gamma^*))^T (\widehat{R}(\theta_0) - \widehat{R}(\theta^*)) + (\text{piso estocástico dependente de } \sigma^2(\gamma^*));$$

3. (**Generalização**) o gap esperado satisfaz

$$\mathbb{E}[R(\theta_T) - \hat{R}(\theta_T)] \lesssim \frac{\eta T}{m} G(\gamma^*)^2, \quad G(\gamma^*) \leq C(\alpha_L(\gamma^*)G_L + \alpha_H(\gamma^*)G_H),$$

logo reduz-se quando  $G_H$  domina e  $\alpha_H(\gamma^*) = (1 - \gamma^*)^{\underline{w}_H}$  é pequeno;

4. (**Robustez adversarial**) o modelo é  $L_{\text{adv}}(\gamma^*)$ -robusto com

$$L_{\text{adv}}(\gamma^*) \leq L_E(\alpha_L(\gamma^*)\|(M_\theta)_L\|_2 + \alpha_H(\gamma^*)\|(M_\theta)_H\|_2),$$

e o raio robusto por margem aumenta proporcionalmente a  $1/L_{\text{adv}}(\gamma^*)$ ;

5. (**Contraprova do alternativo isotrópico**) para um ruído isotrópico que atenua low e high por um mesmo fator, não é possível, em geral, obter simultaneamente (i) supressão preferencial da parcela espúria high, (ii) melhoria do condicionamento local, e (iii) preservação do sinal low; portanto o alternativo falha em reproduzir o regime de  $\gamma^*$ .

*Prova (estrutura).* **Passo 1 (Existência de  $\gamma^*$ ).** Segue do Teorema de benefício condicionado (A) e da otimização de  $\gamma$  (Seção 3), pois a parcela high decai mais rápido que a low:  $\alpha_H(\gamma) \ll \alpha_L(\gamma)$  para pesos separados.

**Passo 2 (Convergência).** Segue do Teorema de convergência sob PL local no landscape suavizado (Seção 2), com constantes dependentes de  $\gamma$  via  $q_P(\gamma)$ .

**Passo 3 (Generalização).** Segue do Teorema de estabilidade (Seção 4) e do fato de que o tamanho efetivo do gradiente contrai seletivamente com  $\gamma$ .

**Passo 4 (Robustez).** Segue do Teorema 10 desta Seção.

**Passo 5 (Contraprova).** Segue do Teorema/contraprova da Seção 3: sem seletividade (mesmo fator para low/high), o trade-off não produz um regime em que a supressão do espúrio high ocorra sem degradar o low na mesma ordem.  $\square$

[Como apresentar isso como “descoberta matemática”] O ponto matemático-chave é que a família  $q_P(\gamma) = (1 - \gamma)^{dw(P)}$  implementa uma **regularização anisotrópica no espaço de modos**, com taxa de contração controlada por um invariante estrutural  $dw(P)$  (peso/light-cone). Essa anisotropia cria um **intervalo não-trivial** de  $\gamma$  onde:

- remove-se preferencialmente coerências espúrias (melhora risco e estabilidade),

- suaviza-se o landscape (melhora constantes PL/suavidade e convergência),
- reduz-se sensibilidade a perturbações (robustez por margem),
- e um alternativo isotrópico não consegue reproduzir o mesmo conjunto de propriedades.

Isso fecha o “ciclo de validação” exigido em banca: *enunciado*  $\rightarrow$  *provas*  $\rightarrow$  *estimadores operacionais*  $\rightarrow$  *contraprova*.

## 12.8 (5.7) Conclusão e extensões (para a redação final do artigo)

**Conclusão.** Demonstramos que ruído quântico seletivo pode ser formalizado como uma regularização anisotrópica que: (i) reduz componentes espúrias (benefício condicionado), (ii) melhora convergência via PL local no landscape suavizado, (iii) reduz gap de generalização via estabilidade, e (iv) aumenta robustez adversarial por contração de norma em modos de alto peso.

### Extensões imediatas.

- Estender a robustez para outras métricas (traço  $\|\cdot\|_1$ ) via bounds de normas cruzadas;
- Ligar  $dw(P)$  explicitamente ao light-cone do circuito e profundidade (bound combinatório por arquitetura);
- Incorporar *noise-aware training* com  $\gamma$  aprendível e garantias de permanência em  $\Gamma_{\text{adm}}$ .