

# **Aprendizagem estatística em altas dimensões**

**Florencia Leonardi**

# Conteúdo

- \* Comparação de modelos
- \* Decomposição do erro esperado “fora da amostra”
- \* Seleção de modelos
- \* Validação cruzada

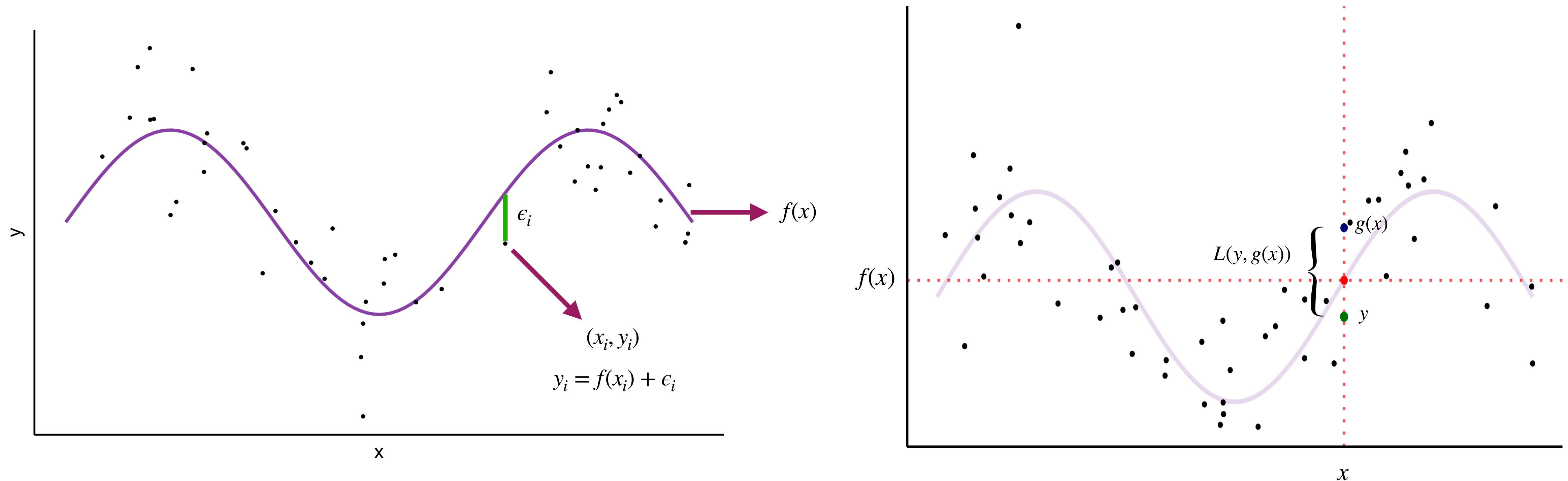
# Objetivos da aprendizagem estatística supervisionada

Formalização do problema de aprendizagem estatística supervisionada:

- \* Uma função objetivo  $f: \mathcal{X} \rightarrow \mathcal{Y}$  desconhecida
- \* Um conjunto de dados (exemplos)  $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$
- \* Uma função de custo  $L: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_{\geq 0}$
- \* Uma família de funções candidatas  $\mathcal{G}$  (modelo)

O objetivo da aprendizagem estatística é “aprender” a função objetivo  $f$  a partir de um conjunto de dados observado  $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$

# Objetivos da aprendizagem estatística supervisionada



# Como escolher $g$ ?

**Objetivo:** escolher  $g \in \mathcal{G}$  que minimize  $\mathbb{E}[L(y, g(x))]$

Lembrando:

$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  é um estimador de  $\mathbb{E}(X)$  se  $x_1, \dots, x_n$  é uma amostra da variável  $X$

**Ideia:** escolher  $g \in \mathcal{G}$  que minimize  $\frac{1}{n} \sum_{i=1}^n L(y_i, g(x_i))$

Esta ideia funciona bem quando a complexidade da família  $\mathcal{G}$  é a adequada para o problema, mas pode ser ruim em vários outros casos !

# Tipos de erro

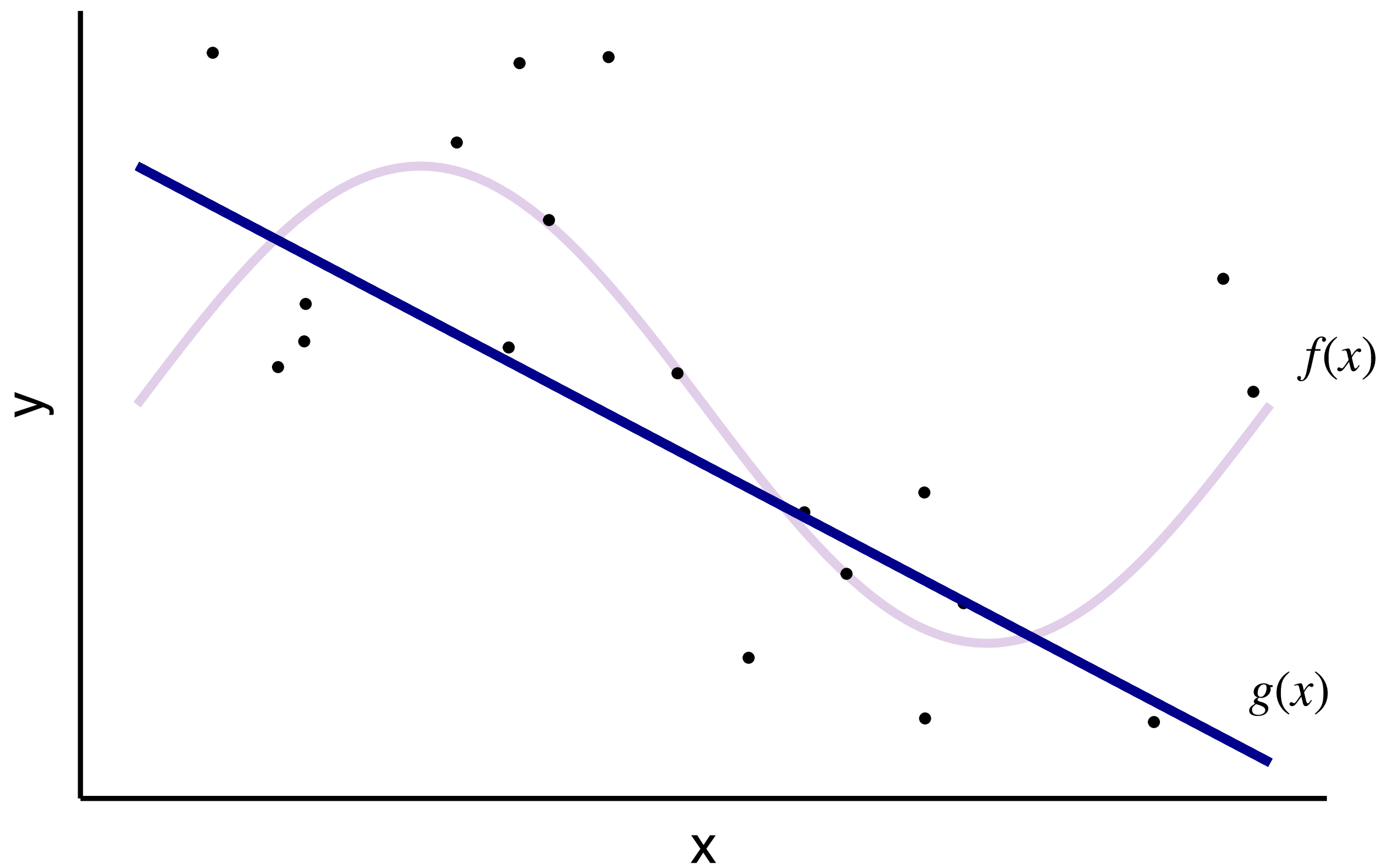
$$\widehat{E}_D(g) = \frac{1}{n} \sum_{i=1}^n L(y_i, g(x_i)) \quad \longrightarrow \quad \text{Erro estimado "dentro da amostra"}$$

$$E_D(g) = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n L(y_i, g(x_i)) \right] \quad \longrightarrow \quad \text{Erro esperado "dentro da amostra"}$$

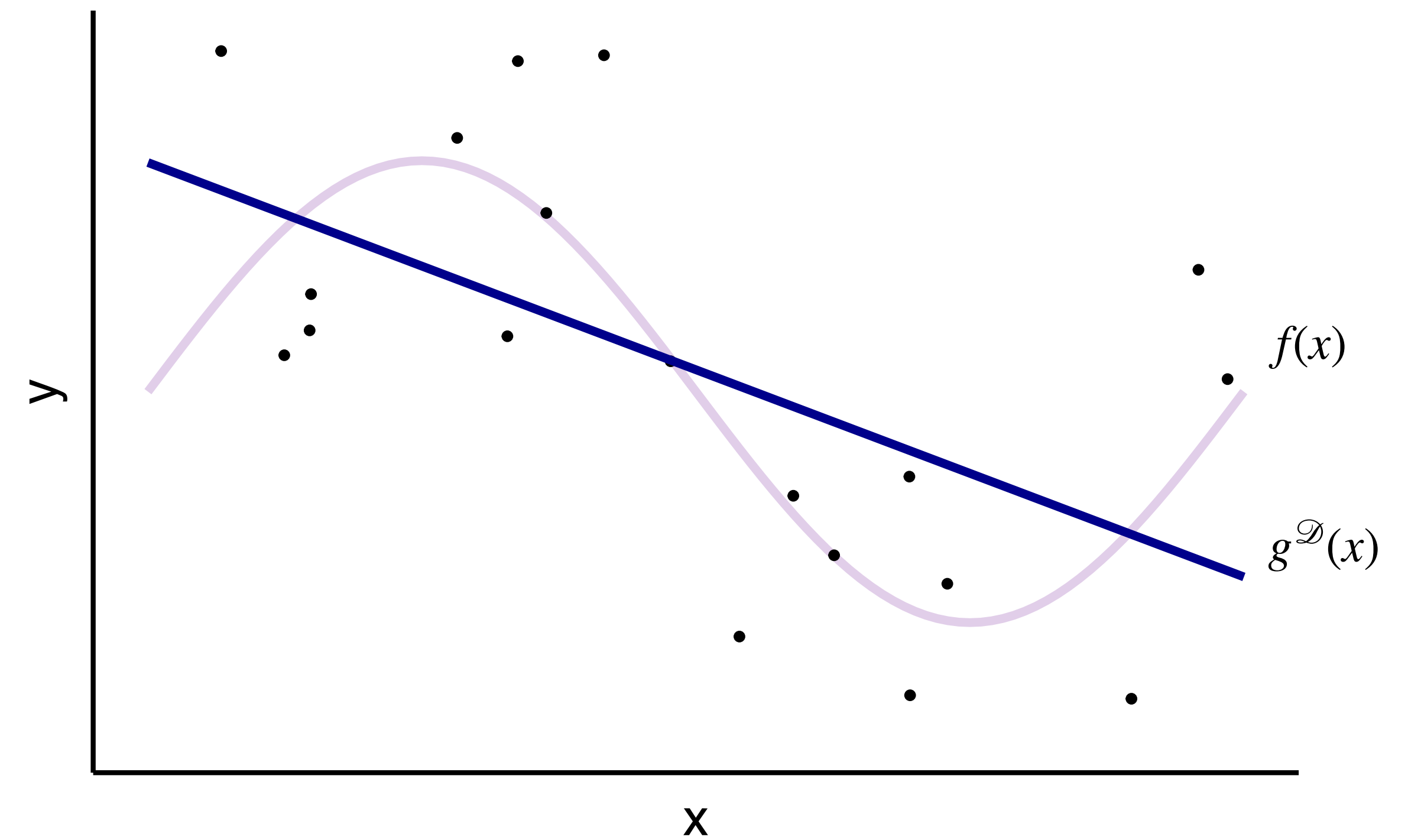
$$E_F(g) = \mathbb{E}(L(y, g(x))) \quad \longrightarrow \quad \text{Erro esperado "fora da amostra"}$$

$$\widehat{E}_F(g) ? \quad \longrightarrow \quad \text{Erro estimado "fora da amostra"}$$

# A função $g$ que nos interessa depende dos dados!

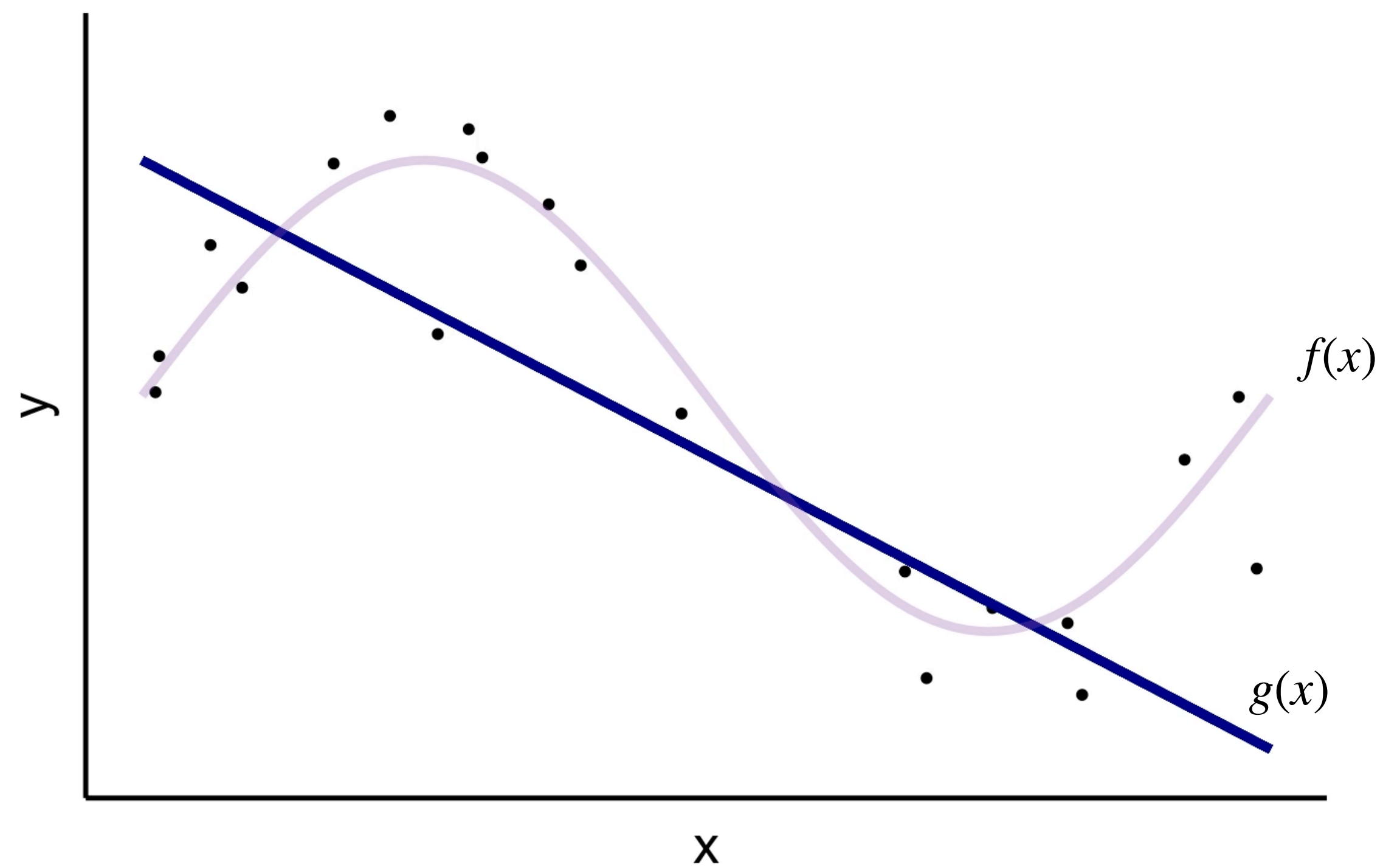


Função  $g$  fixa

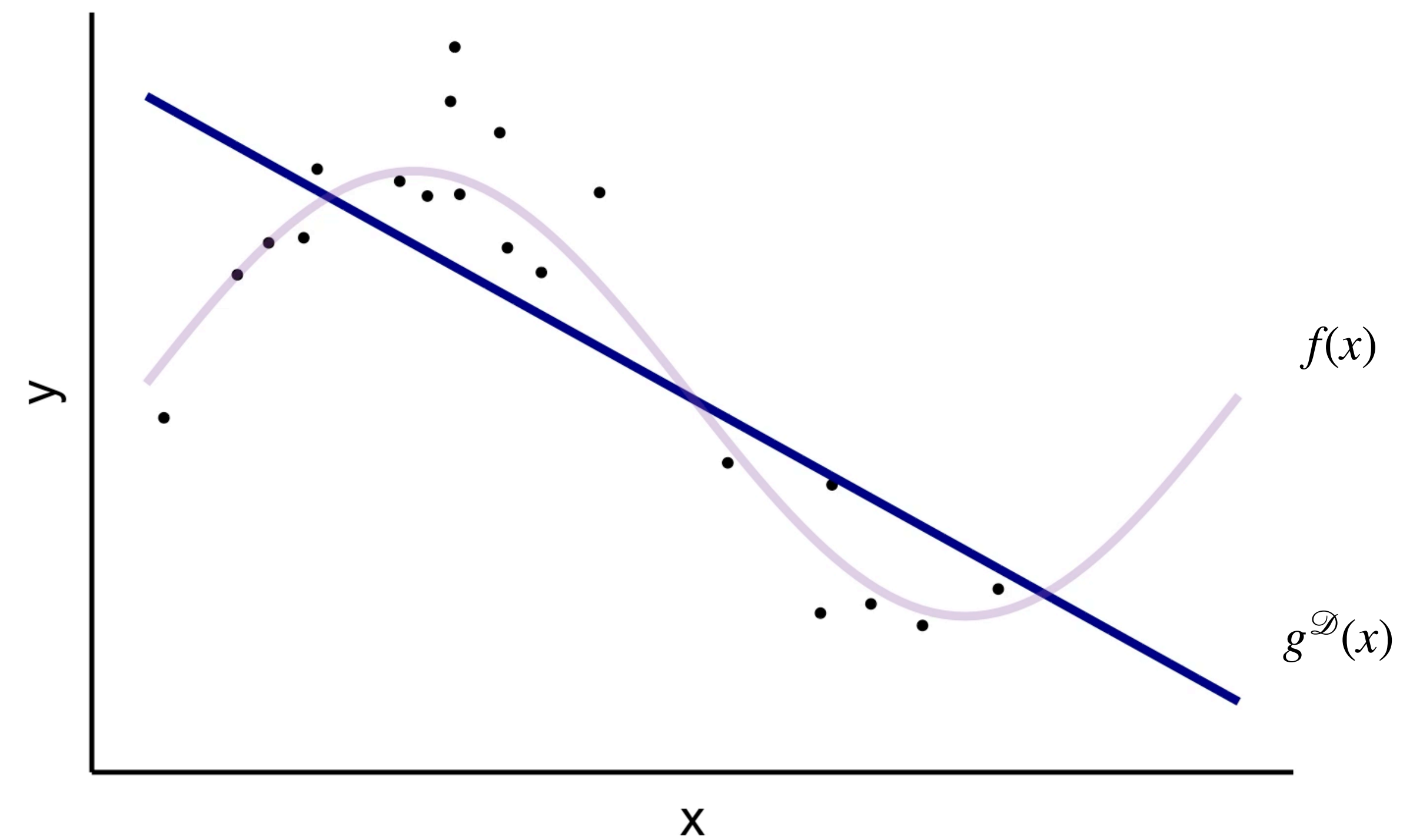


Função  $g$  que minimiza  $\widehat{E}_D(g)$

A função  $g$  que nos interessa depende dos dados!



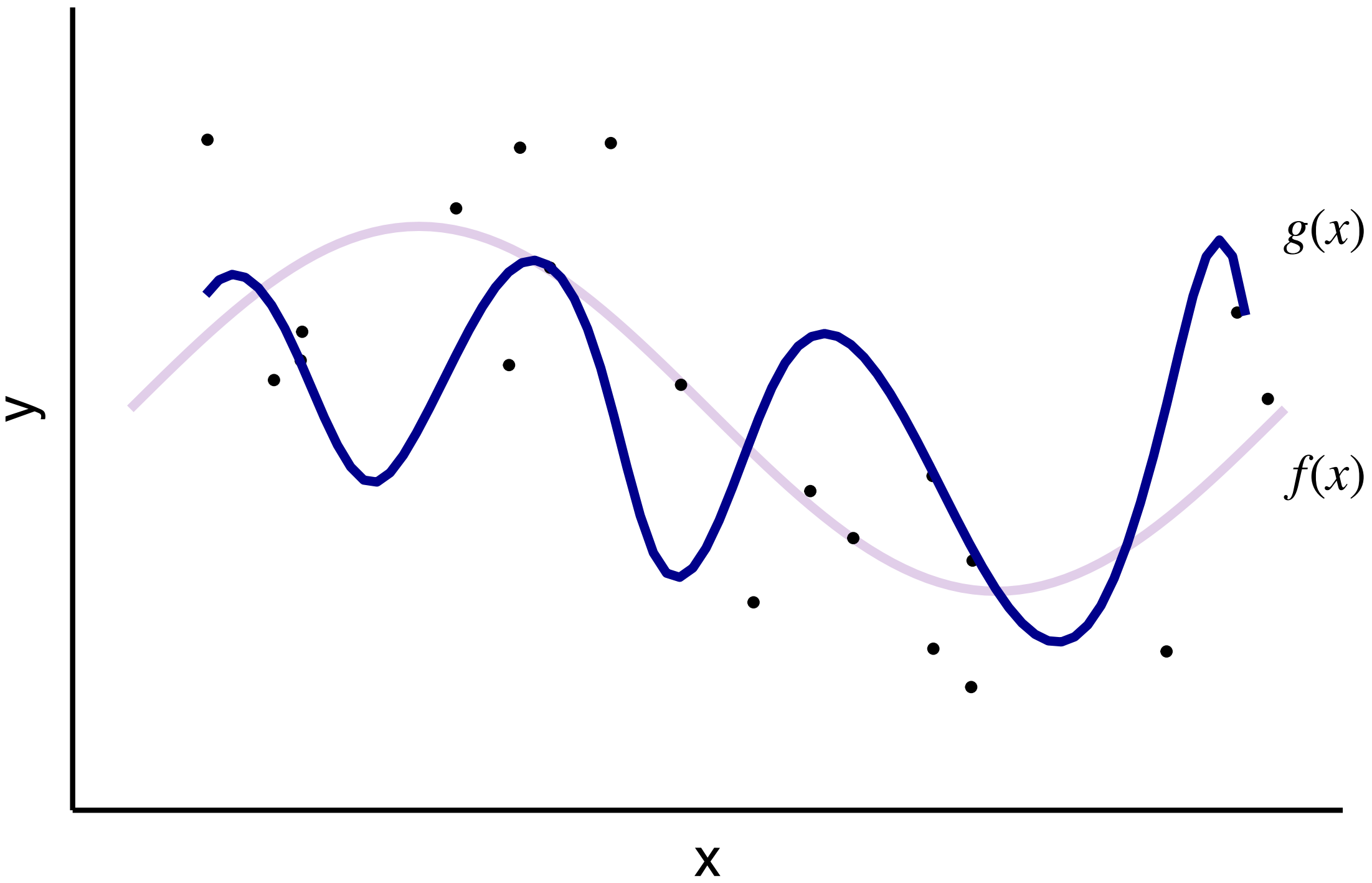
Função  $g$  fixa



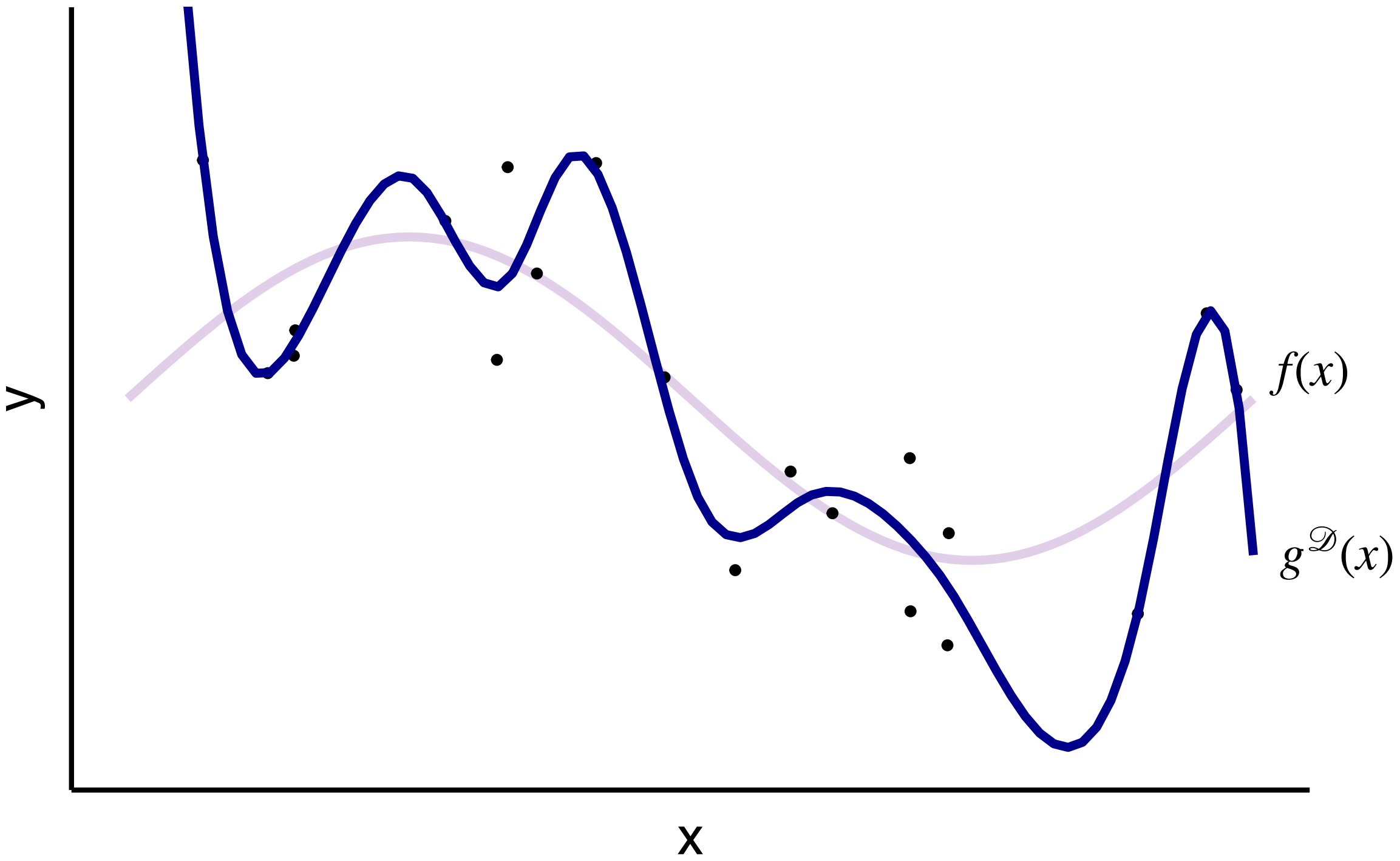
Função  $g$  que minimiza  $\widehat{E}_D(g)$



A função  $g$  que nos interessa depende dos dados!

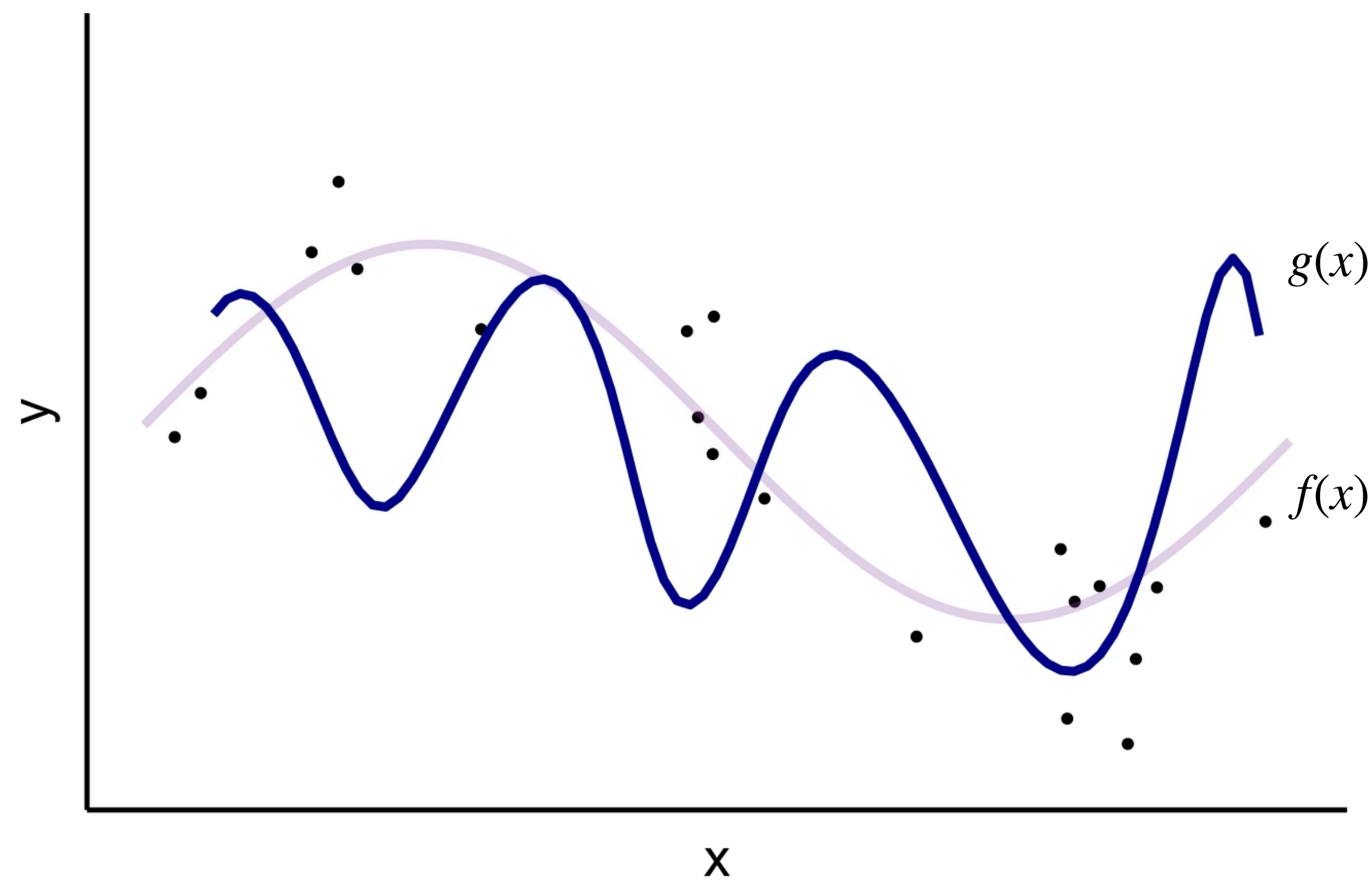


Função  $g$  fixa

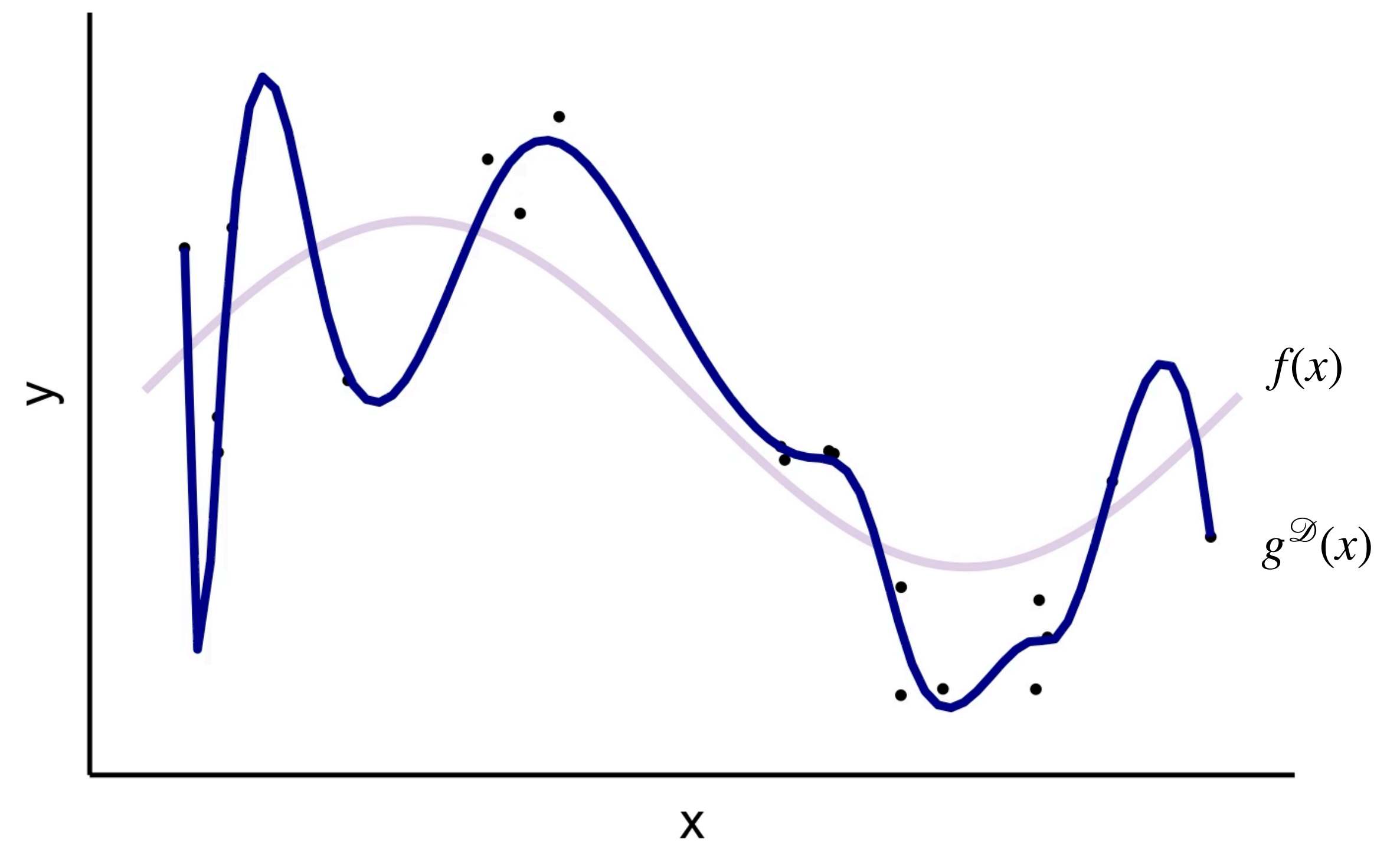


Função  $g$  que minimiza  $\widehat{E}_D(g)$

A função  $g$  que nos interessa depende dos dados!

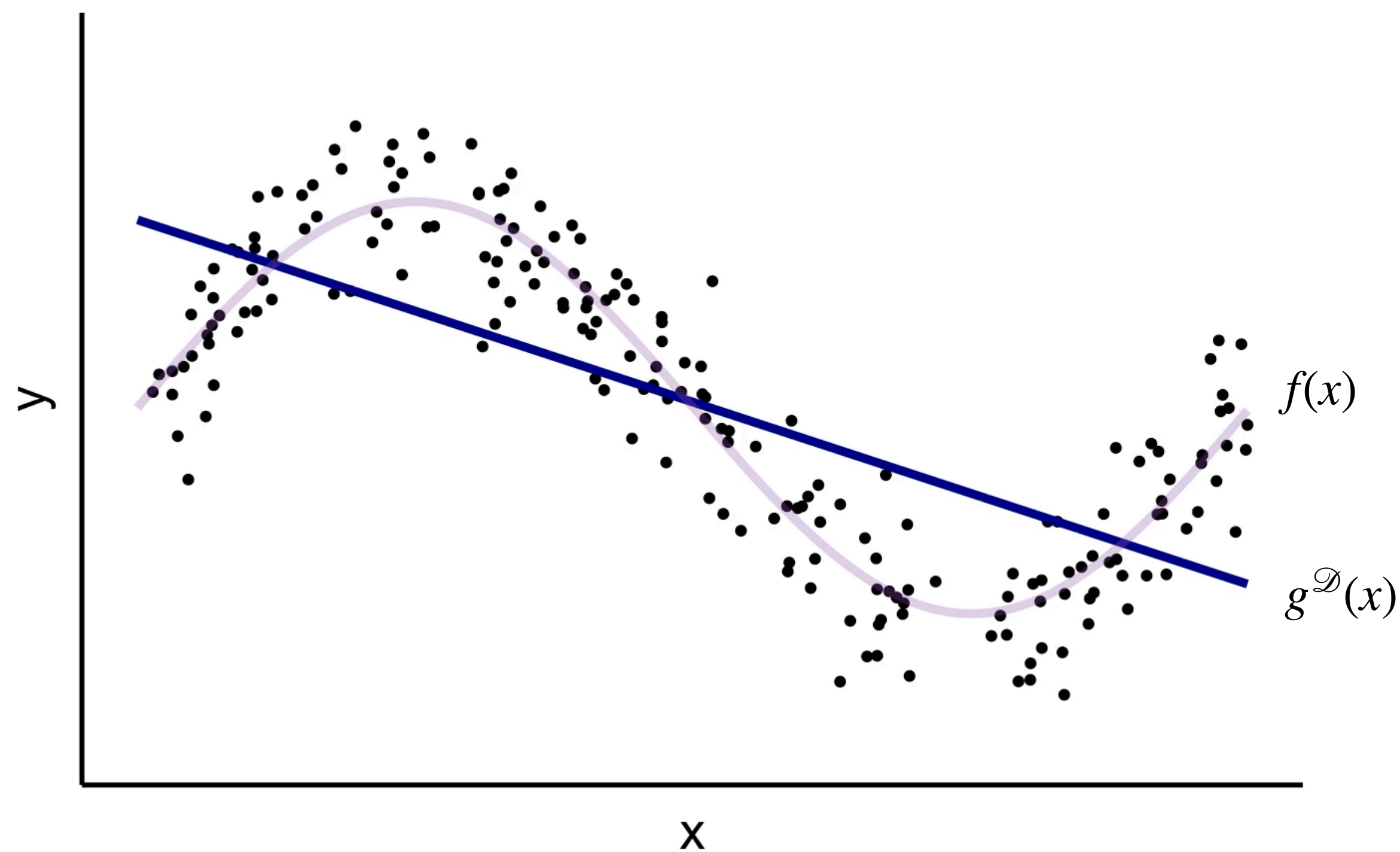


Função  $g$  fixa

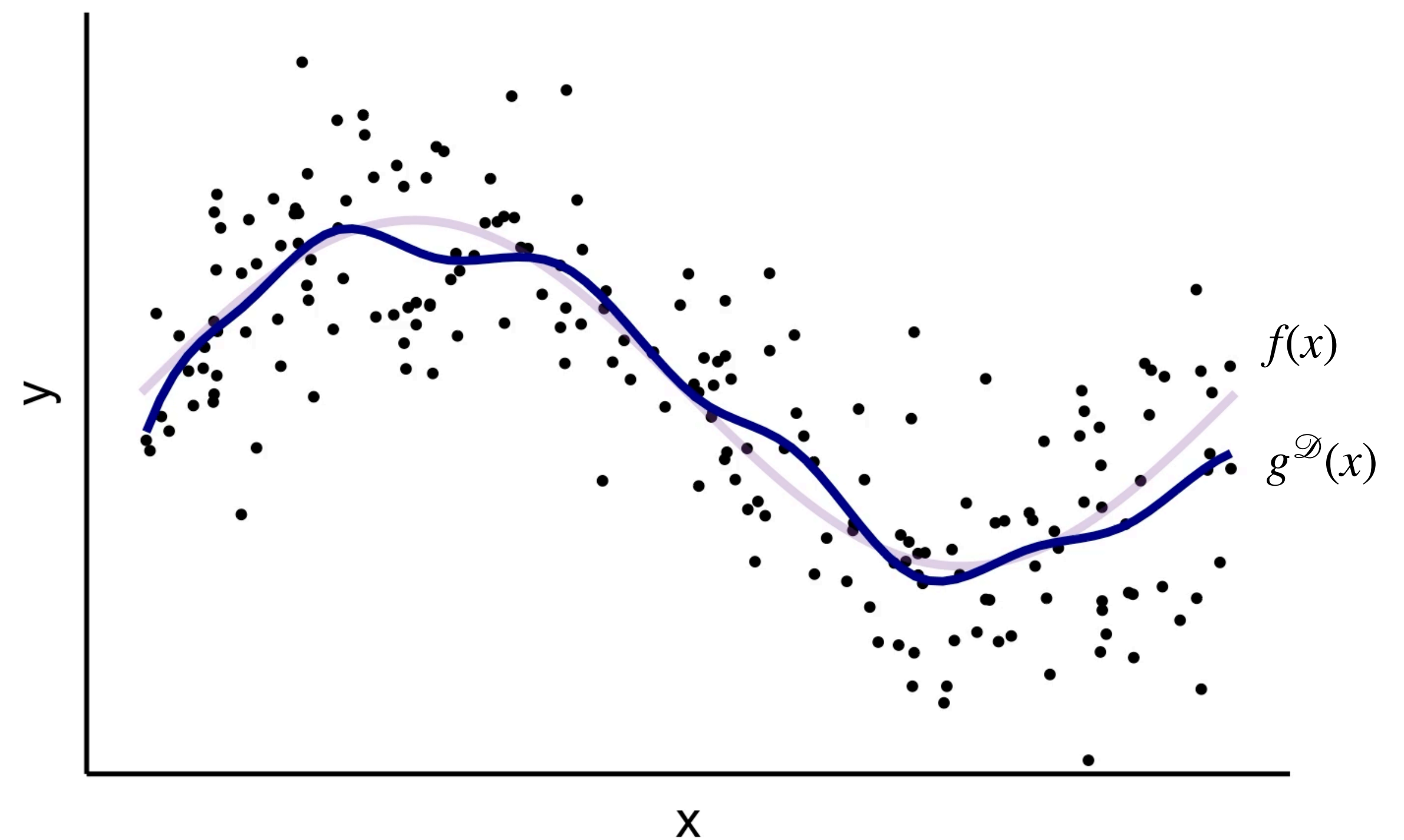


Função  $g$  que minimiza  $\widehat{E}_D(g)$

## E se o tamanho da amostra for maior ... ?



Função  $g$  que minimiza  $\widehat{E}_D(g)$   
 $\mathcal{G}$  = funções lineares



Função  $g$  que minimiza  $\widehat{E}_D(g)$   
 $\mathcal{G}$  = modelo de splines de grau 12

# Decomposição do erro de predição

Para cada  $g \in \mathcal{G}$  nós definimos  $E_F(g) = \mathbb{E}[L(y, g(x))]$

Mas nós queremos avaliar o erro esperado fora da amostra da função  $g^{\mathcal{D}}$ , escolhida minimizando  $\hat{E}_D(g)$  !

Ou seja nós gostaríamos de saber quanto vale  $E_F(g^{\mathcal{D}}) = \mathbb{E}(L(y, g^{\mathcal{D}}(x)))$

A esperança é calculada em relação a uma distribuição de probabilidade.... qual neste caso?

- \* A distribuição dos dados  $\mathcal{D}$  que são i.i.d com distribuição  $p(x, y)$
- \* A distribuição da observação de teste  $(x, y)$  que também tem distribuição  $p(x, y)$  e é independente de  $\mathcal{D}$

# Decomposição do erro de predição

Consideremos a função de custo quadrática  $L(y, \hat{y}) = (y - \hat{y})^2$  no caso de regressão

Dada  $g \in \mathcal{G}$  podemos escrever:

$$\begin{aligned} E_F(g) &= \mathbb{E}[(f(x) + \epsilon - g(x))^2] = \mathbb{E}[(f(x) - g(x))^2 + 2(f(x) - g(x))\epsilon + \epsilon^2] \\ &= \mathbb{E}[(f(x) - g(x))^2] + \mathbb{E}[2(f(x) - g(x))\epsilon] + \mathbb{E}[\epsilon^2] \end{aligned}$$

Como assumimos que as variáveis  $X$  e  $\epsilon$  são independentes e ainda  $\mathbb{E}[\epsilon] = 0$ , podemos verificar que

$$\mathbb{E}(2(f(x) - g(x))\epsilon) = 2\mathbb{E}[f(x) - g(x)]\mathbb{E}[\epsilon] = 0$$

Na formulação do modelo ainda assumimos que  $\mathbb{E}[\epsilon^2] = \sigma^2$ , logo obtemos que

$$E_F(g) = \mathbb{E}[(f(x) - g(x))^2] + \sigma^2$$

# Decomposição do erro de predição

$$E_F(g) = \mathbb{E}[(f(x) - g(x))^2] + \sigma^2$$

erro redutível  
(classe  $\mathcal{G}$  + algoritmo)

erro irreduzível  
(ruído)

# Decomposição do erro de predição

Consideremos agora o erro redutível

$$\mathbb{E}[(f(x) - g^{\mathcal{D}}(x))^2] = \mathbb{E}_{(x,y)} \mathbb{E}_{\mathcal{D}}[(f(x) - g^{\mathcal{D}}(x))^2]$$

Para um  $x$  fixo, podemos pensar em  $g^{\mathcal{D}}(x)$  como um estimador de  $f(x)$

Aqui,  $g^{\mathcal{D}}$  é uma função obtida por um método qualquer, não necessariamente a que minimiza  $\widehat{E}_D(g)$

Lembrando:

$$\text{EQM}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \text{Viés}(\hat{\theta})^2 + \text{Var}(\hat{\theta})$$

Logo:

$$\mathbb{E}_{\mathcal{D}}[(f(x) - g^{\mathcal{D}}(x))^2] = \text{Viés}(g^{\mathcal{D}}(x))^2 + \text{Var}(g^{\mathcal{D}}(x))$$



# Decomposição do erro de predição

$$\mathbb{E}_{\mathcal{D}}[(f(x) - g^{\mathcal{D}}(x))^2] = \text{Viés}(g^{\mathcal{D}}(x))^2 + \text{Var}(g^{\mathcal{D}}(x))$$

$$\text{Viés}(g^{\mathcal{D}}(x))^2 = [\mathbb{E}_{\mathcal{D}}(g^{\mathcal{D}}(x)) - f(x)]^2$$

$$\text{Var}(g^{\mathcal{D}}(x)) = \mathbb{E}_{\mathcal{D}}[(g^{\mathcal{D}}(x) - \underbrace{\mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(x)]}_{\bar{g}(x)})^2]$$

Muitas vezes  $\bar{g}(x)$  é a melhor aproximação de  $f$  dentro da classe  $\mathcal{G}$

$$g^* = \inf_{g \in \mathcal{G}} \mathbb{E}_x[(f(x) - g(x))^2]$$



# Exemplo

$$f: [-1,1] \rightarrow \mathbb{R}$$

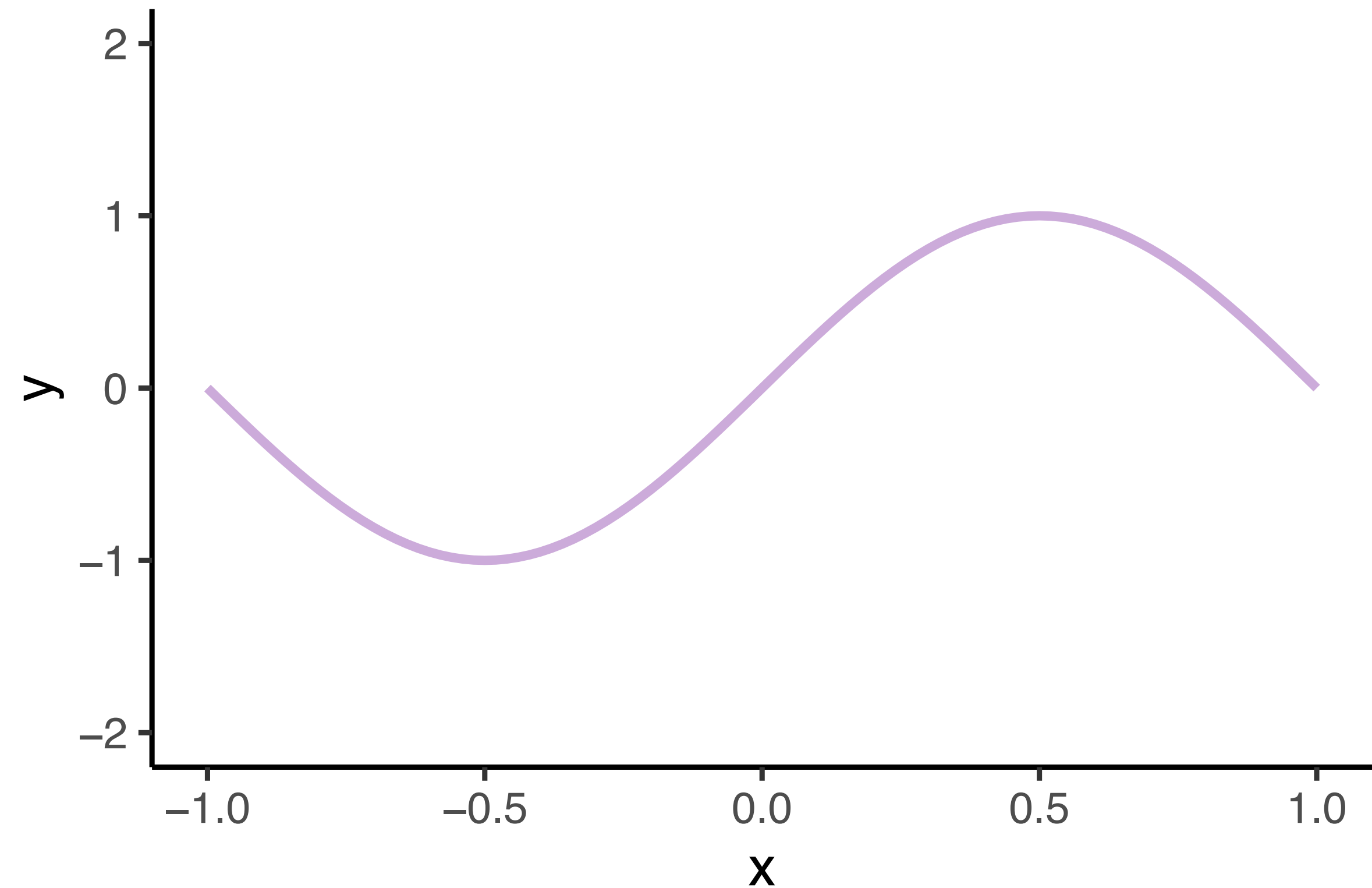
$$f(x) = \sin(\pi x)$$

$$\epsilon = 0 \quad y = f(x)$$

Duas classes de modelos

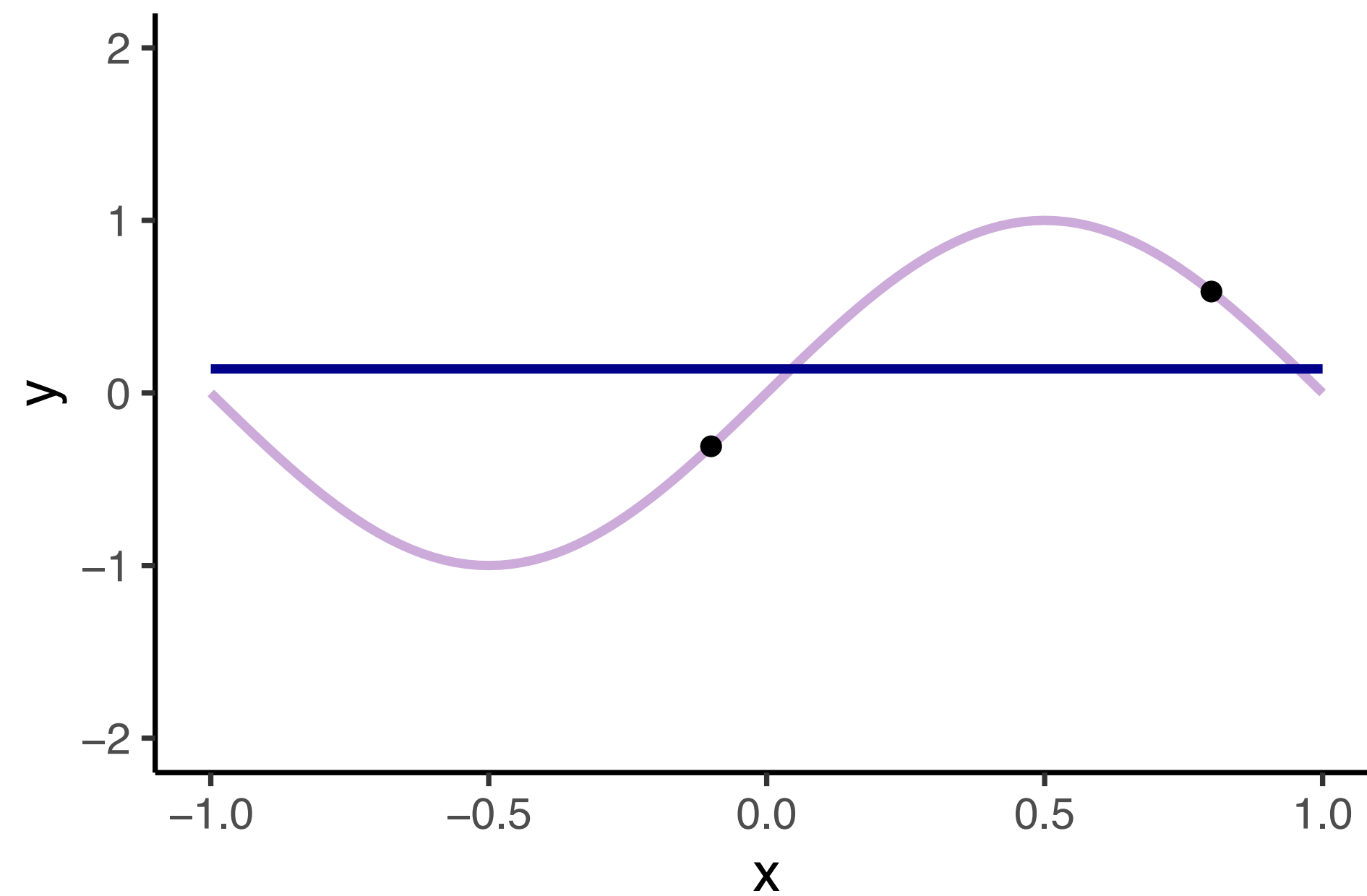
$$\mathcal{G}_1 = \{g(x) = \beta_0: \beta_0 \in \mathbb{R}\}$$

$$\mathcal{G}_2 = \{g(x) = \beta_0 + \beta_1 x: (\beta_0, \beta_1) \in \mathbb{R}^2\}$$

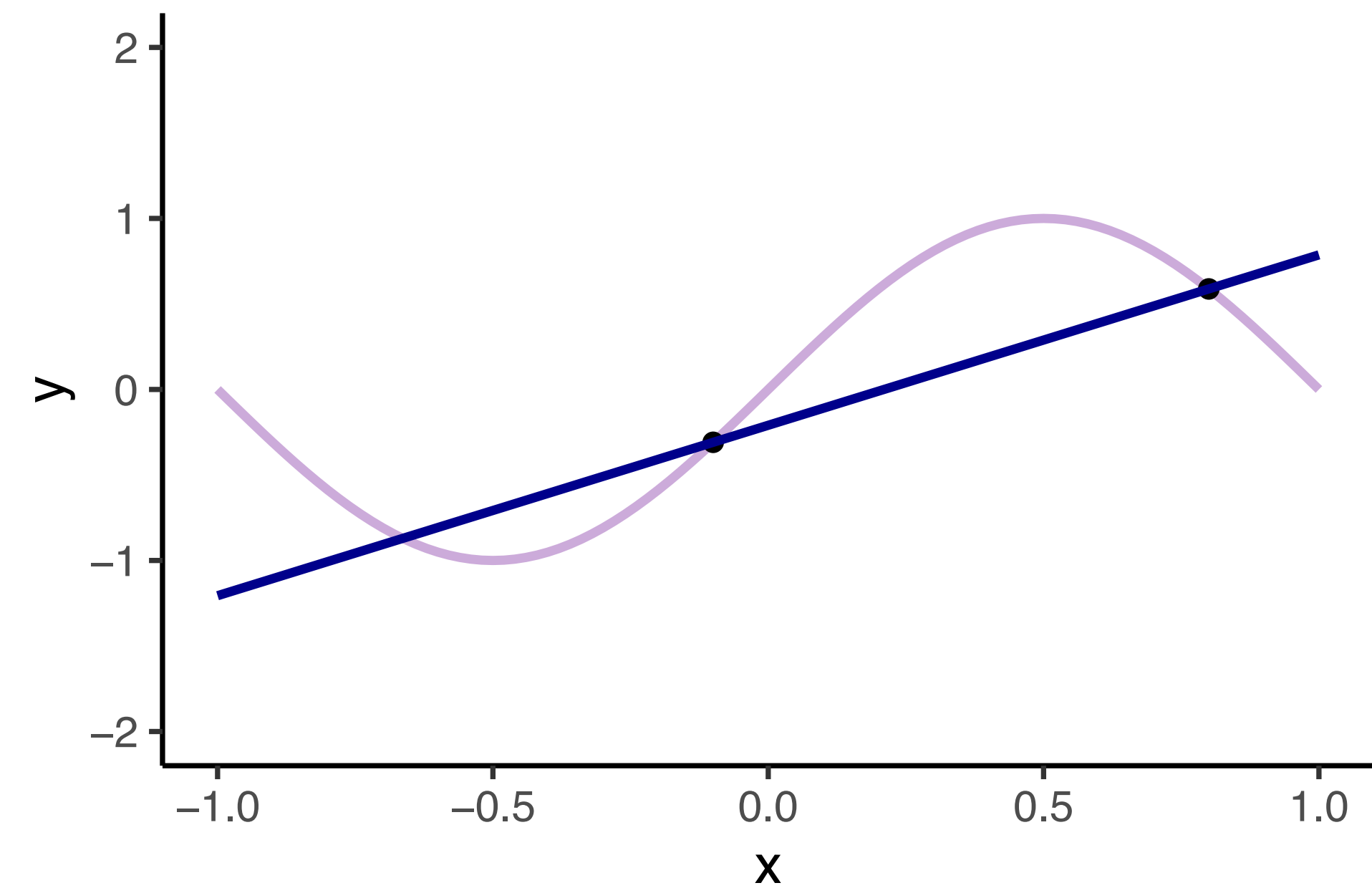


# Exemplo

Suponhamos que nosso conjunto de dados  $\mathcal{D}$  só tem dois pontos:



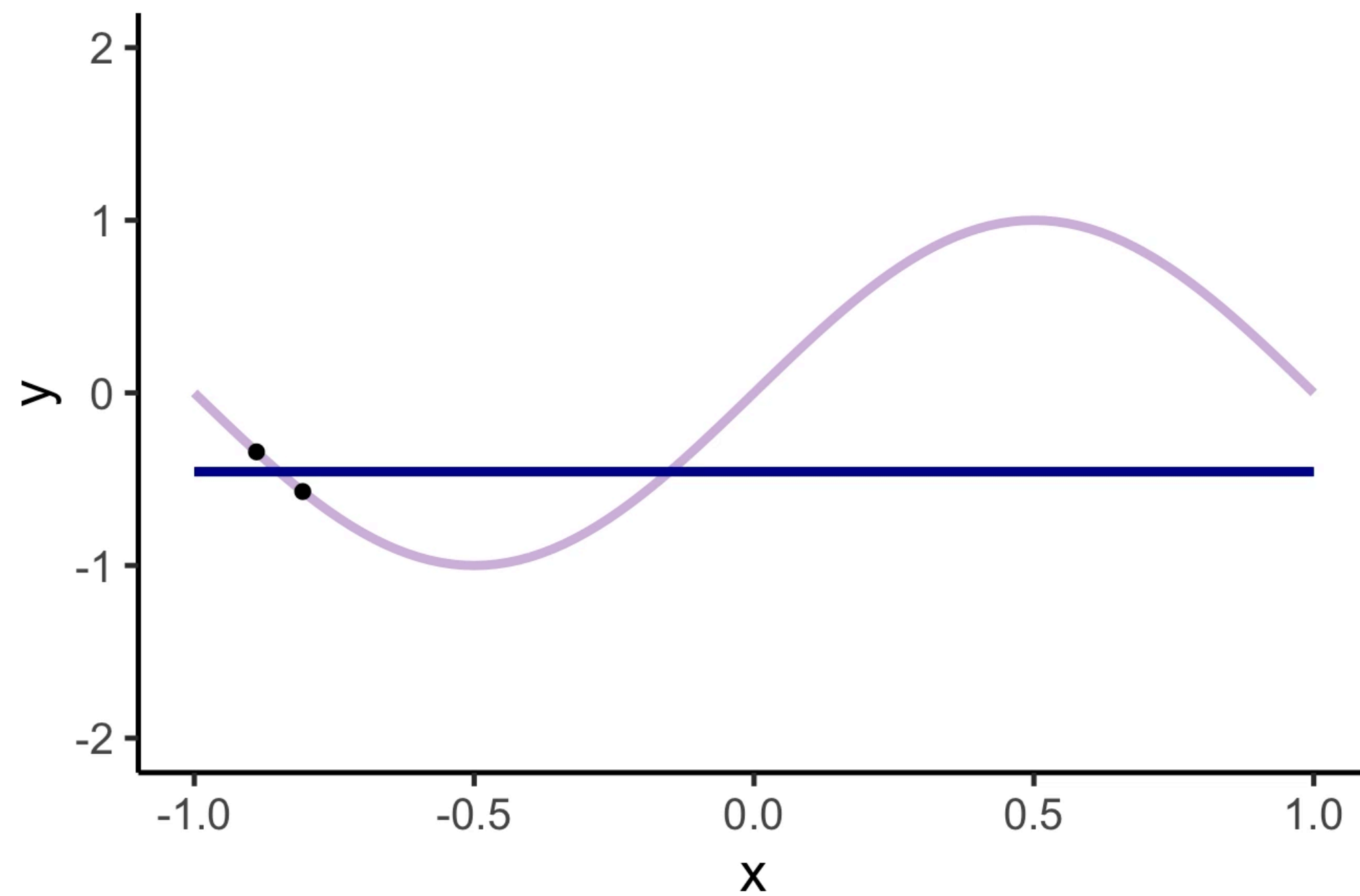
$$\mathcal{G}_1 = \{g(x) = \beta_0 : \beta_0 \in \mathbb{R}\}$$



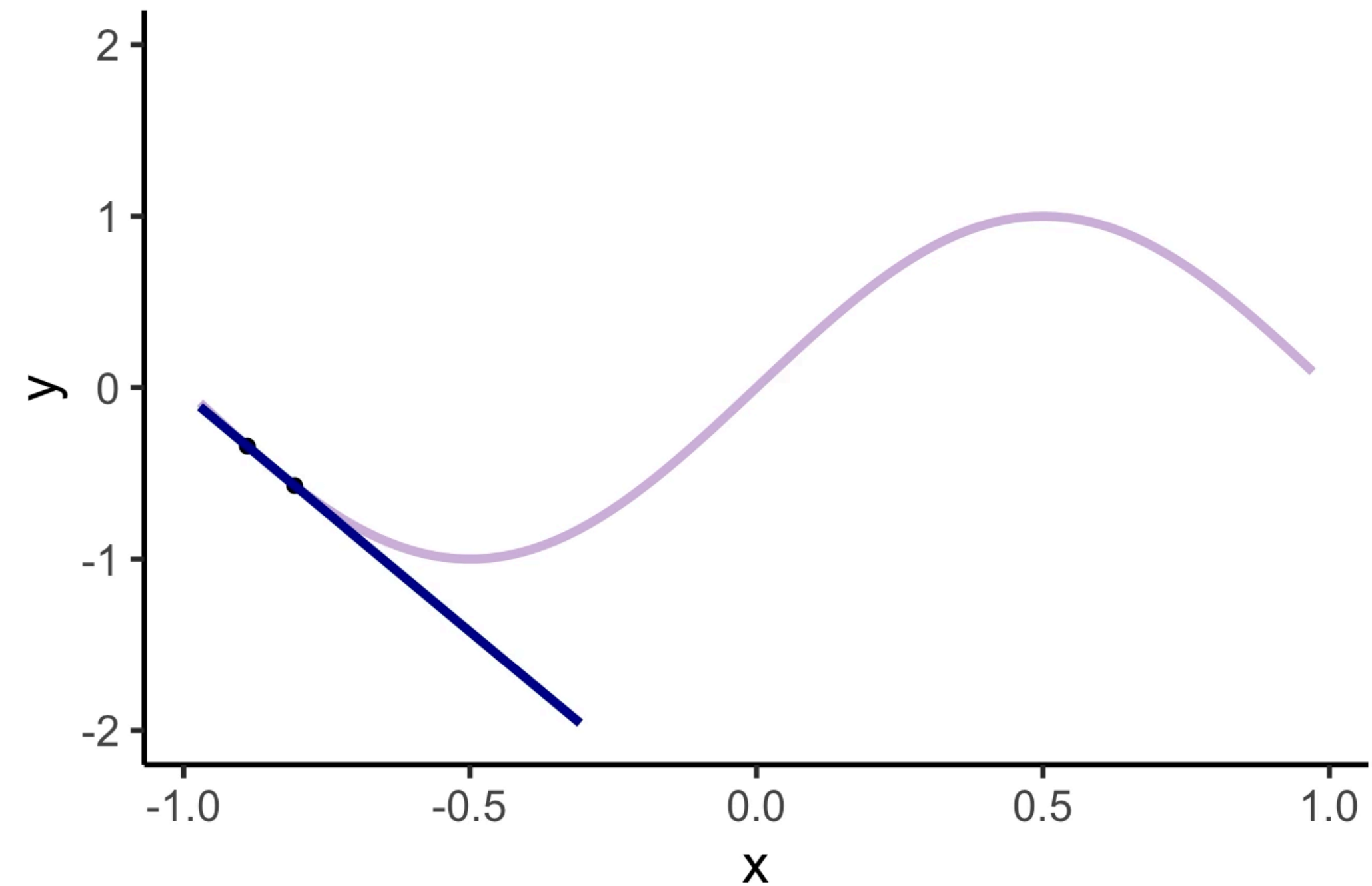
$$\mathcal{G}_2 = \{g(x) = \beta_0 + \beta_1 x : (\beta_0, \beta_1) \in \mathbb{R}^2\}$$

# Exemplo

Suponhamos que nosso conjunto de dados  $\mathcal{D}$  só tem dois pontos:

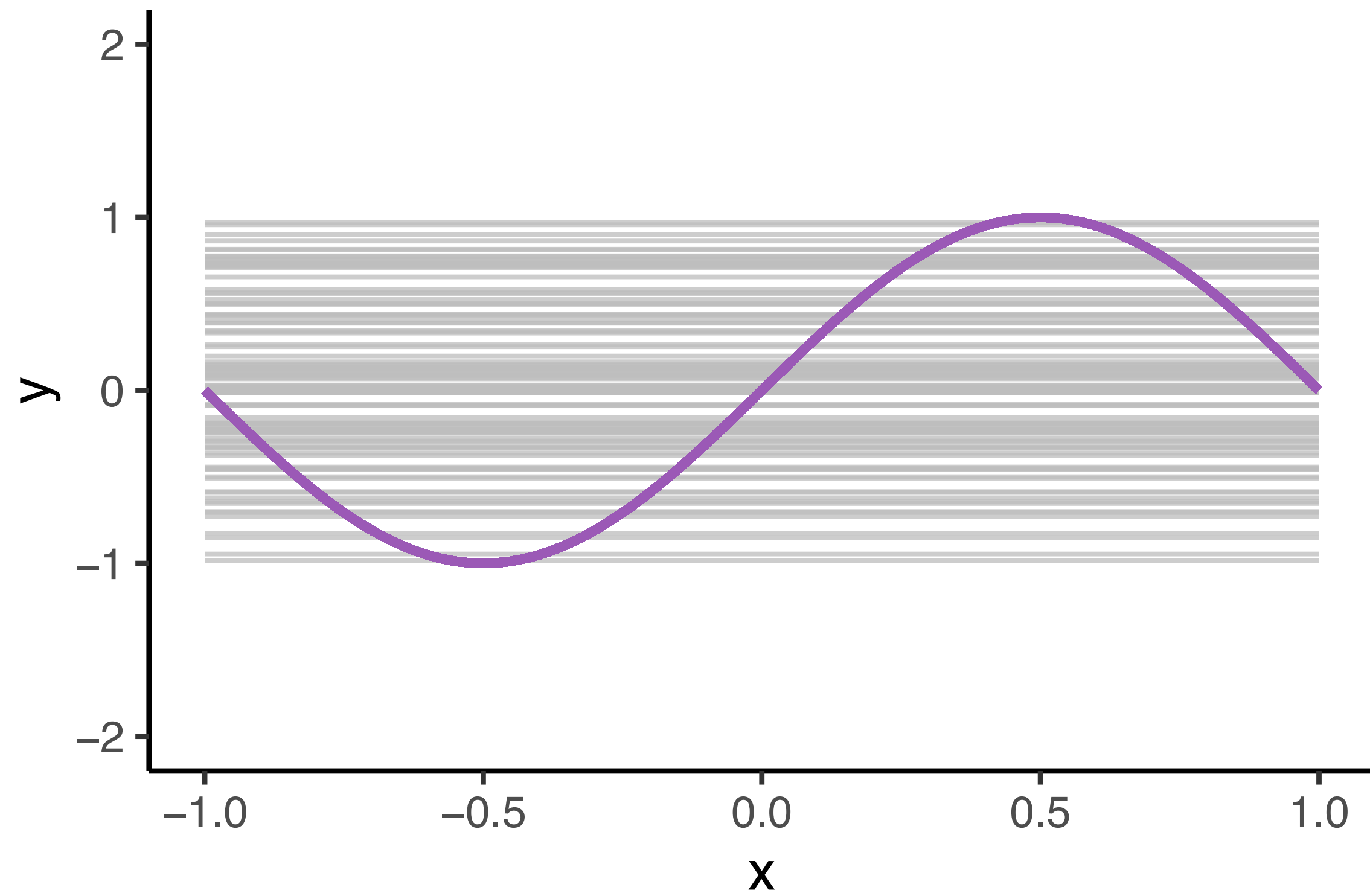


$$\mathcal{G}_1 = \{g(x) = \beta_0 : \beta_0 \in \mathbb{R}\}$$

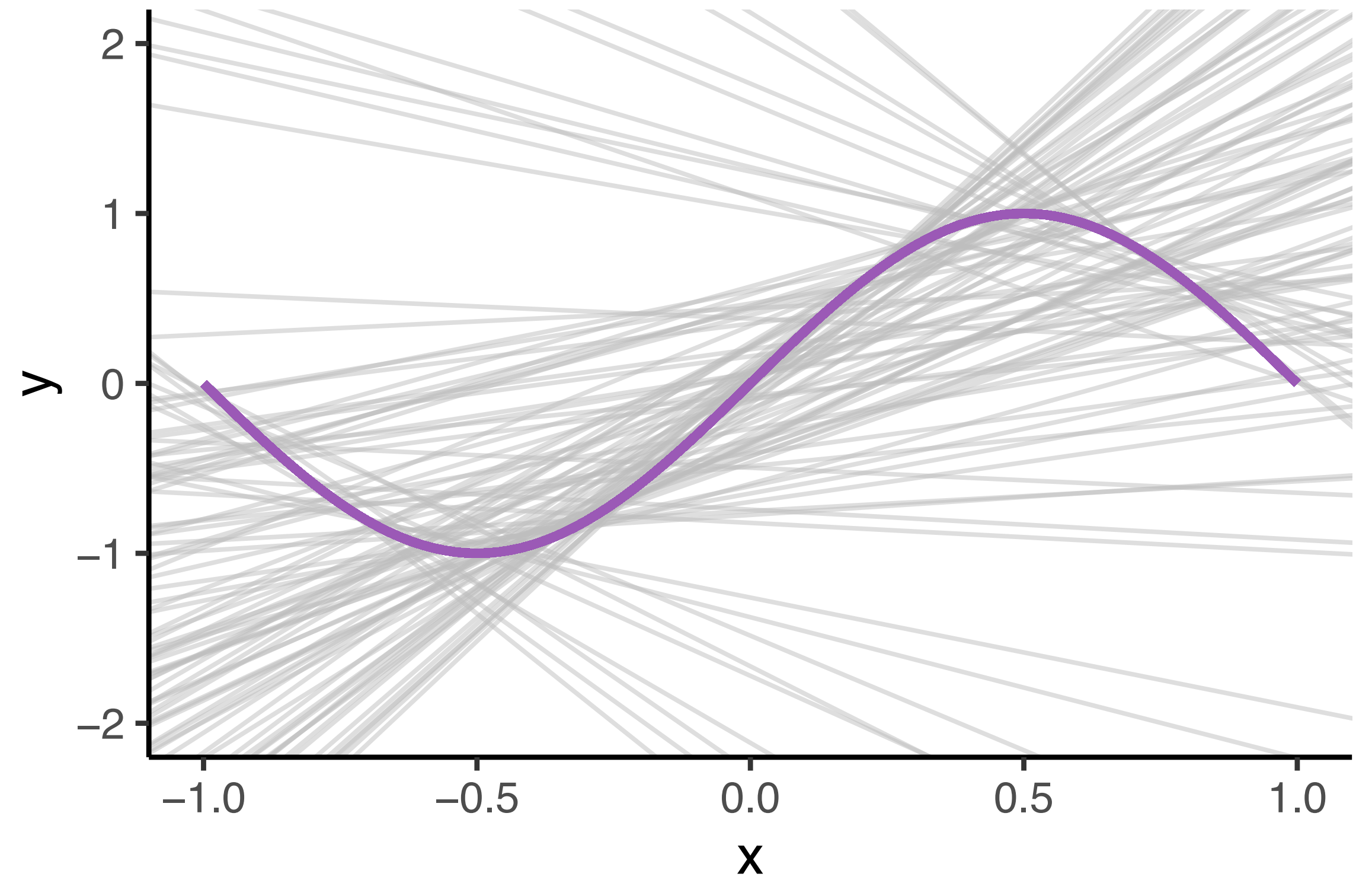


$$\mathcal{G}_2 = \{g(x) = \beta_0 + \beta_1 x : (\beta_0, \beta_1) \in \mathbb{R}^2\}$$

# Exemplo

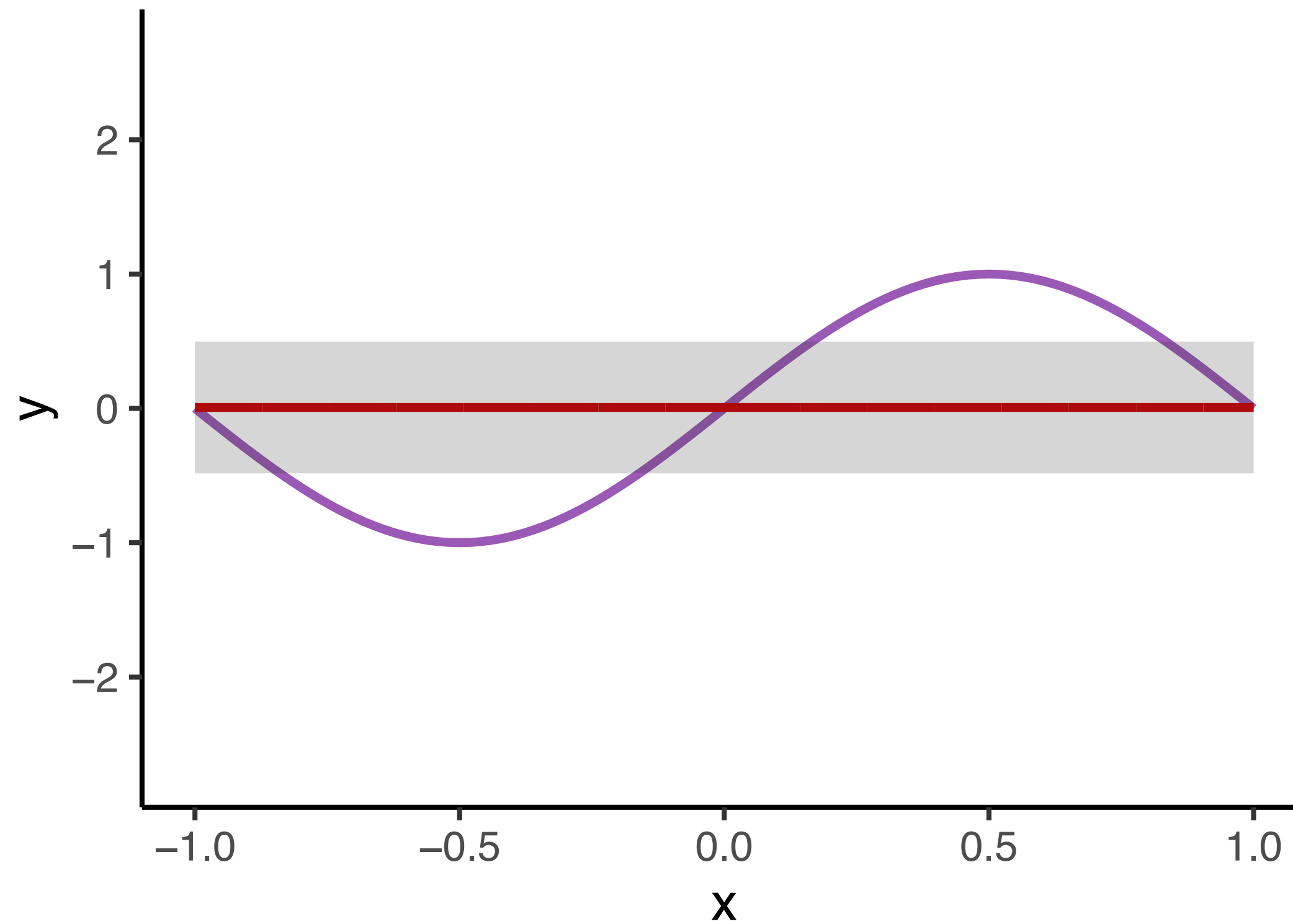


$$\mathcal{G}_1 = \{g(x) = \beta_0 : \beta_0 \in \mathbb{R}\}$$



$$\mathcal{G}_2 = \{g(x) = \beta_0 + \beta_1 x : (\beta_0, \beta_1) \in \mathbb{R}^2\}$$

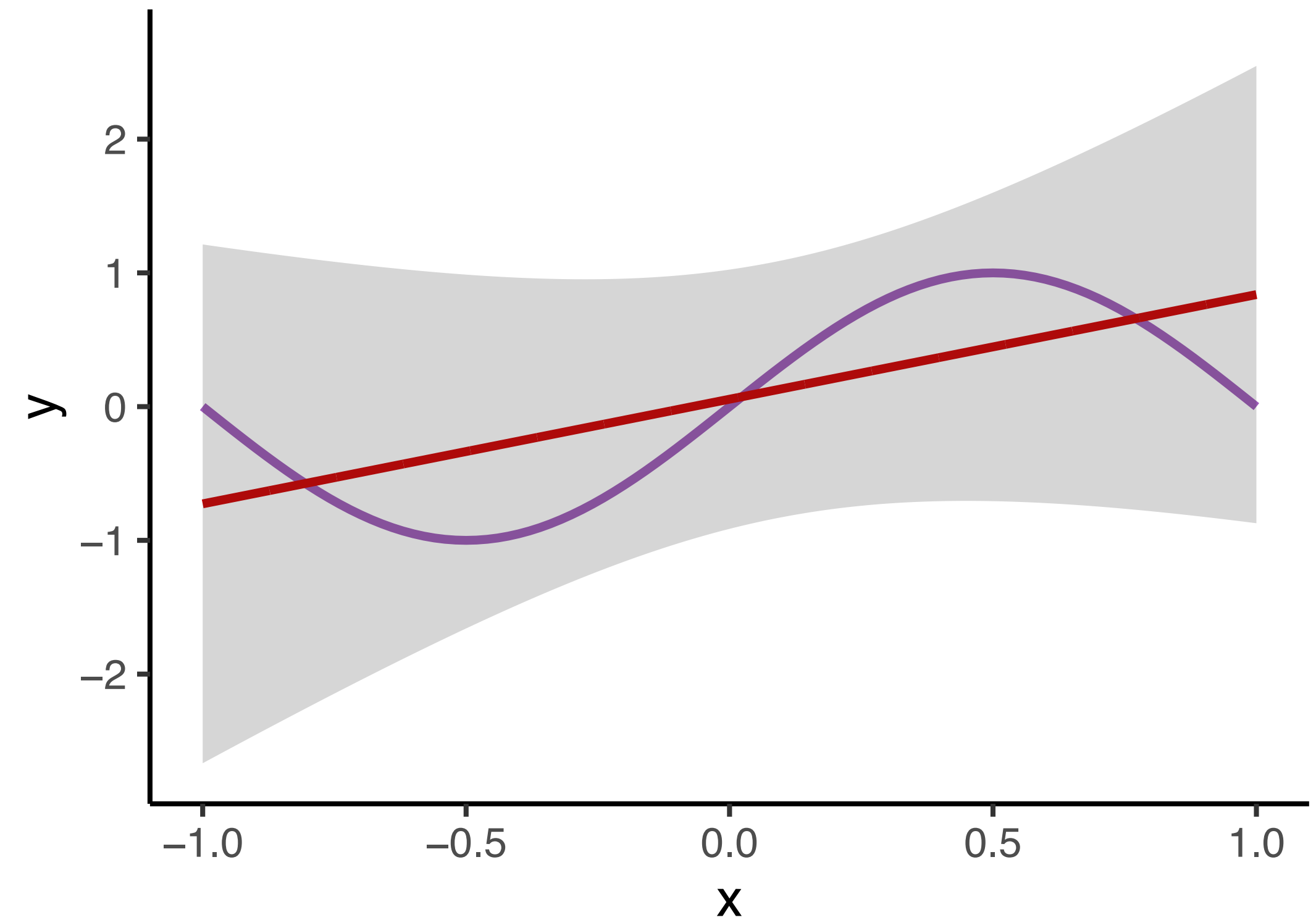
# Exemplo



$$\mathcal{G}_1 = \{g(x) = \beta_0 : \beta_0 \in \mathbb{R}\}$$

$$\text{Viés}^2 = 0,50$$

$$\text{Variância} = 0,25$$

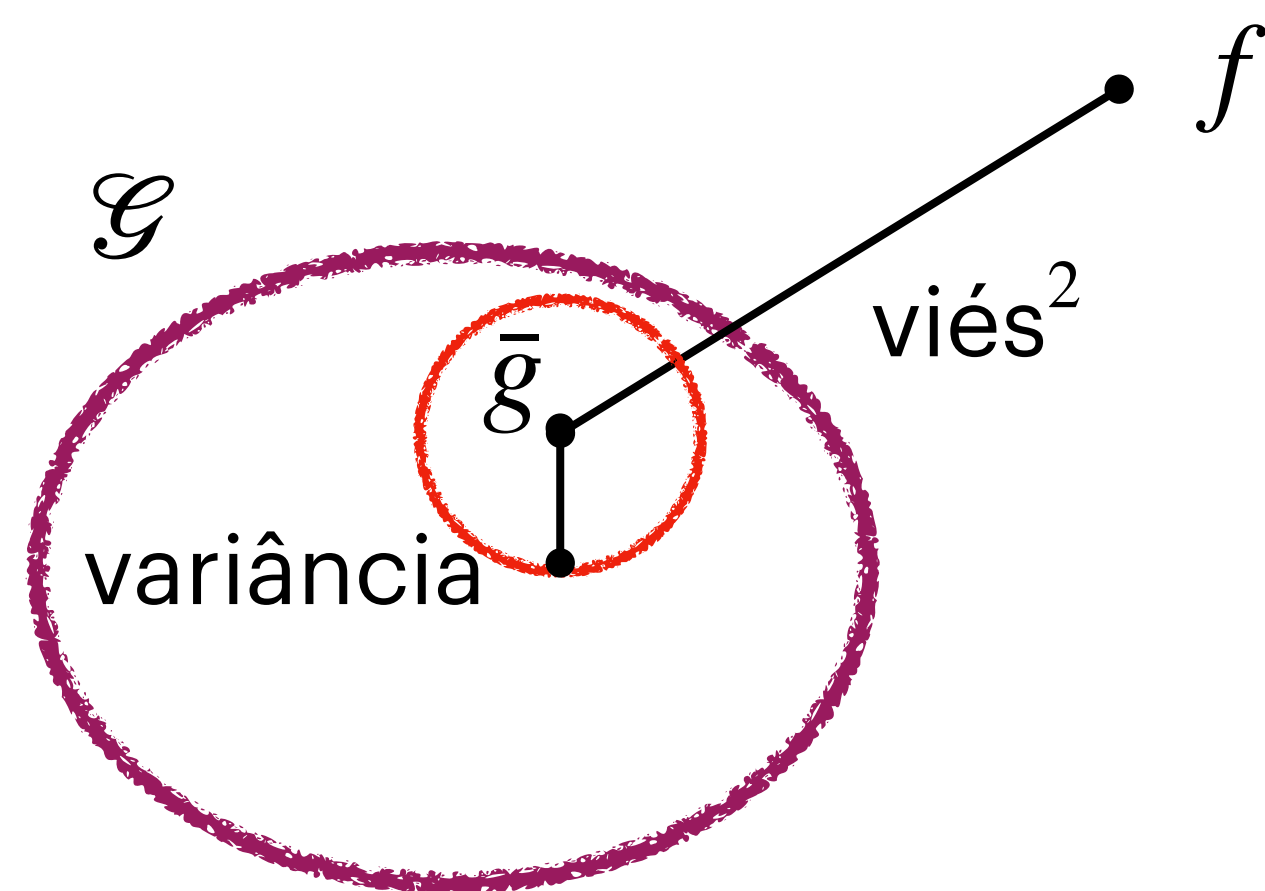


$$\mathcal{G}_2 = \{g(x) = \beta_0 + \beta_1 x : (\beta_0, \beta_1) \in \mathbb{R}^2\}$$

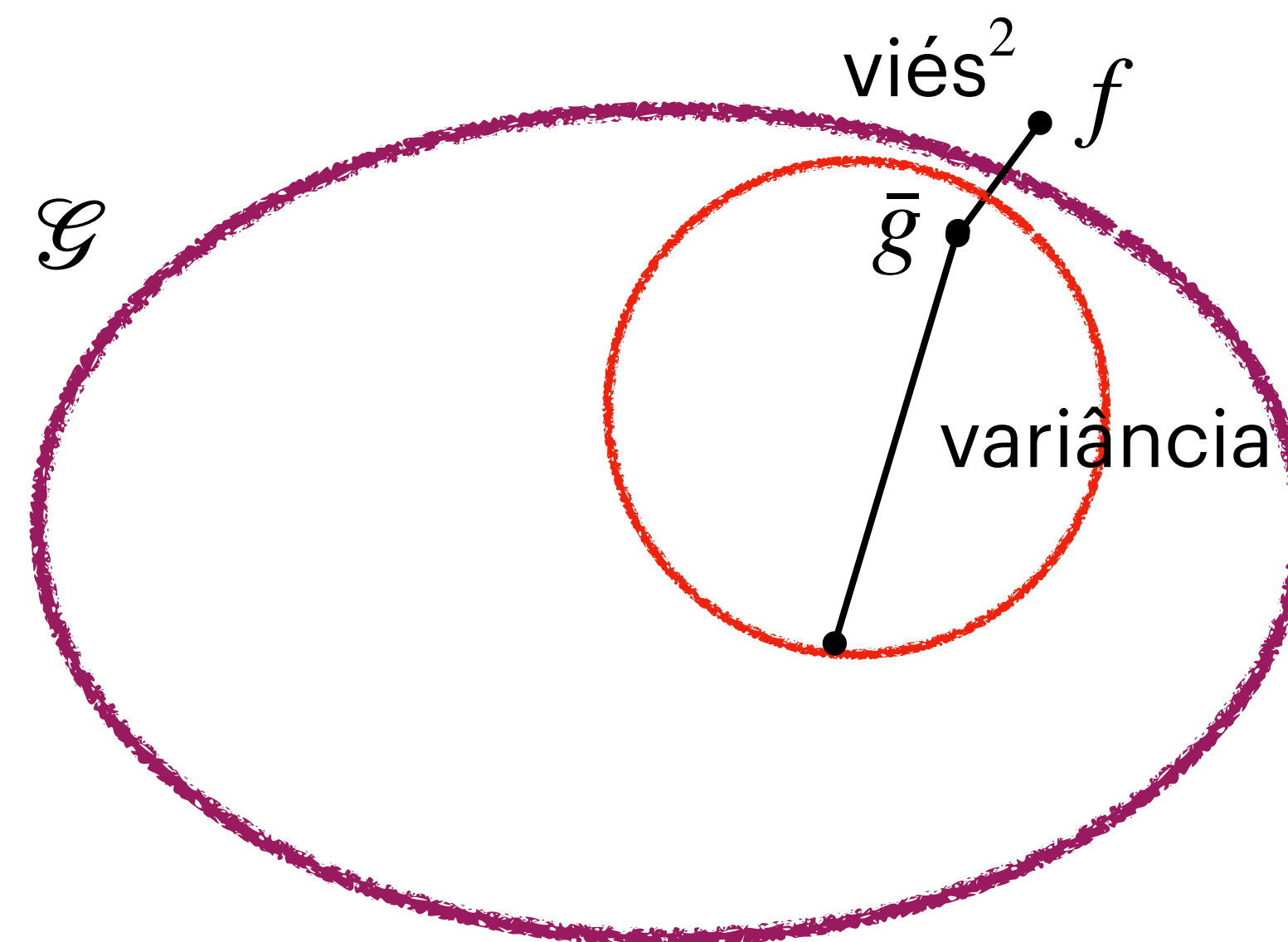
$$\text{Viés}^2 = 0,21$$

$$\text{Variância} = 1,69$$

# O custo benefício

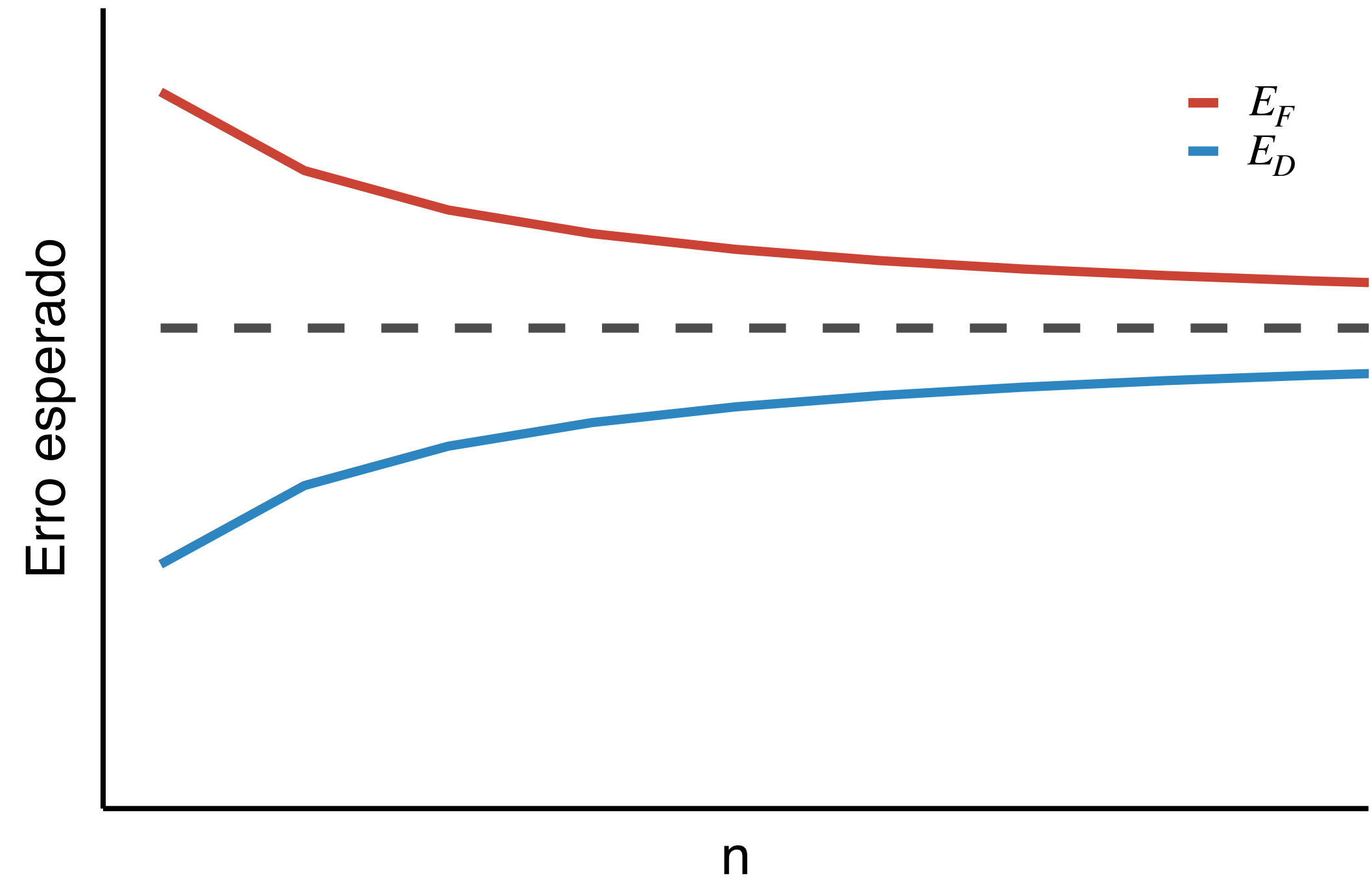


Modelo "simples"

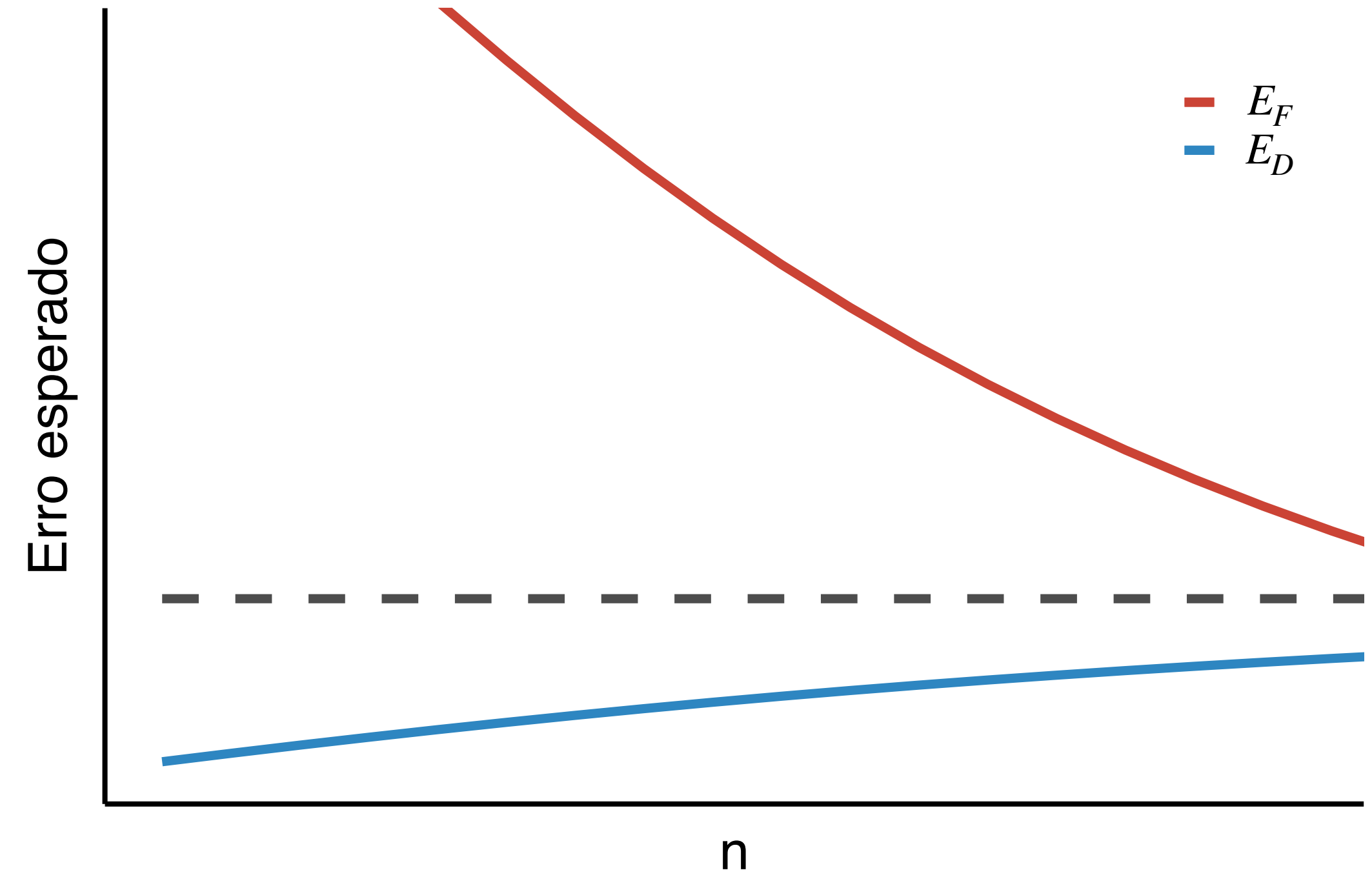


Modelo "complexo"

# Curvas de erro

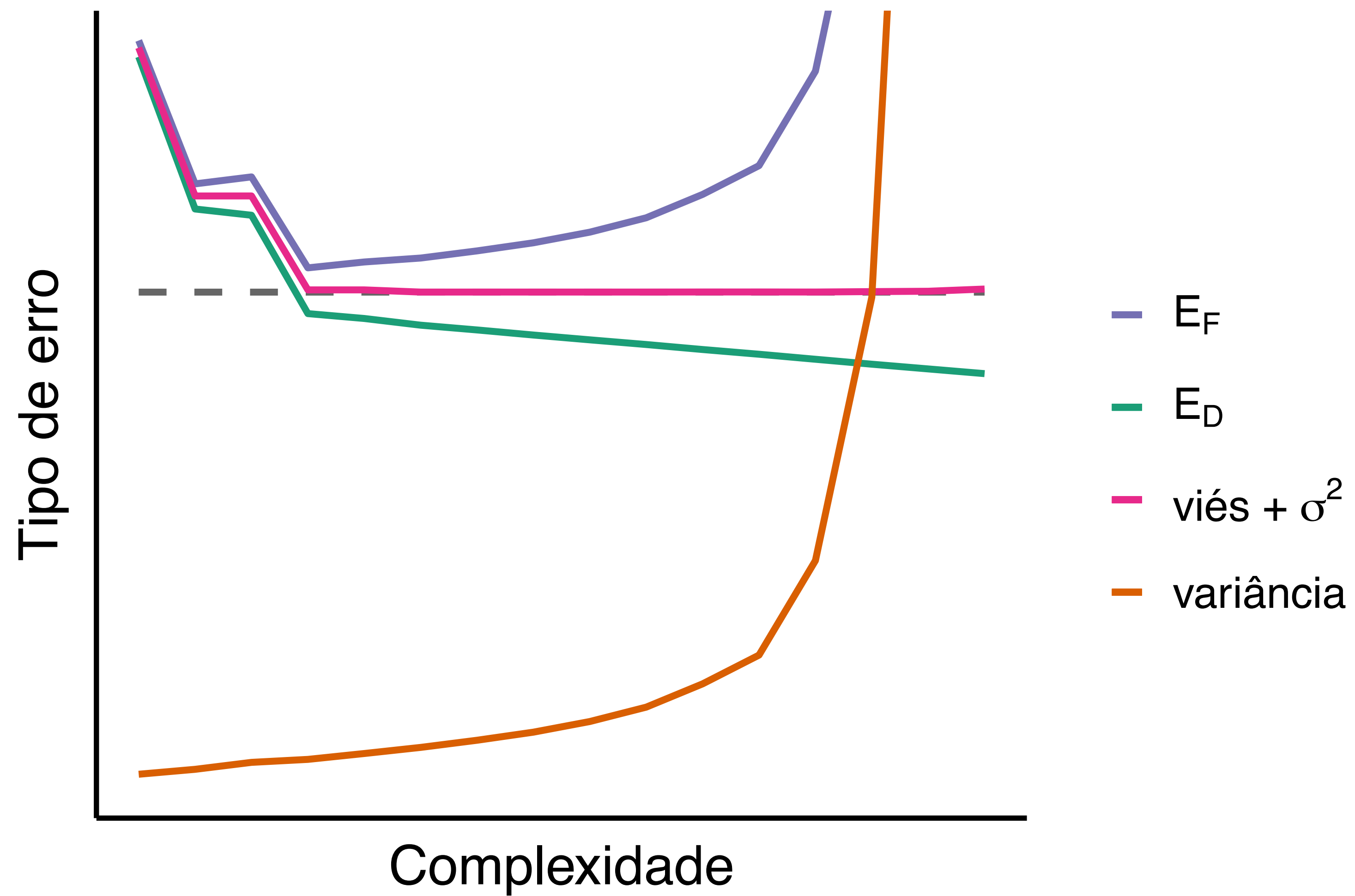


Modelo "simples"



Modelo "complexo"

# Curvas de erro





# O custo benefício

*“A complexidade do modelo deve estar de acordo com os dados disponíveis, e não com a complexidade da função objetivo”*

*Yaser Abu Mostafa - Learning from data - MIT Online Course*

## Estimação do erro “fora da amostra”

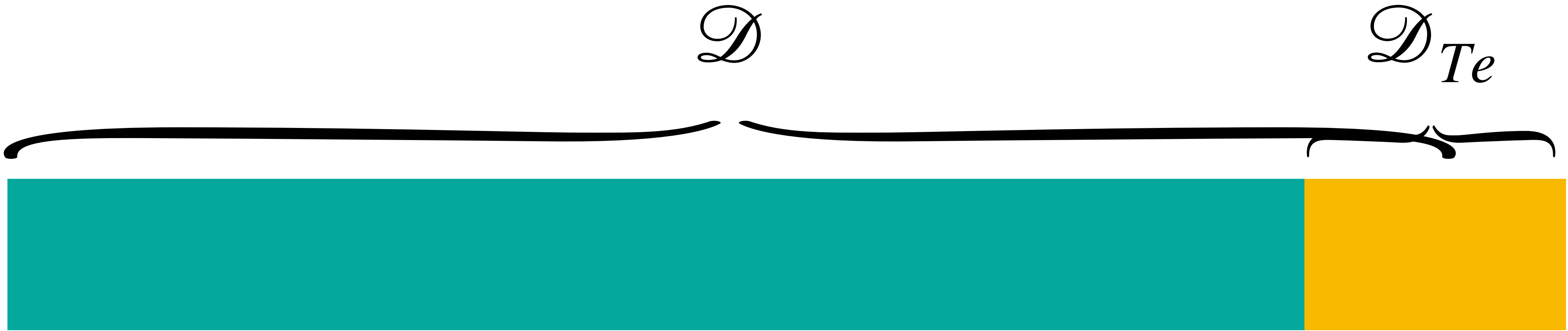
Se nós tivermos acesso a uma amostra independente de  $\mathcal{D}$ , que denotamos por  $\mathcal{D}_{Te} = \{(x_1, y_1), \dots, (x_K, y_K)\}$ , podemos estimar

$$E_F(g) = \mathbb{E}(L(y, g)) \text{ pela média empírica } \hat{E}_F(g) = \frac{1}{K} \sum_{k=1}^K L(y_i, g(x_i))$$

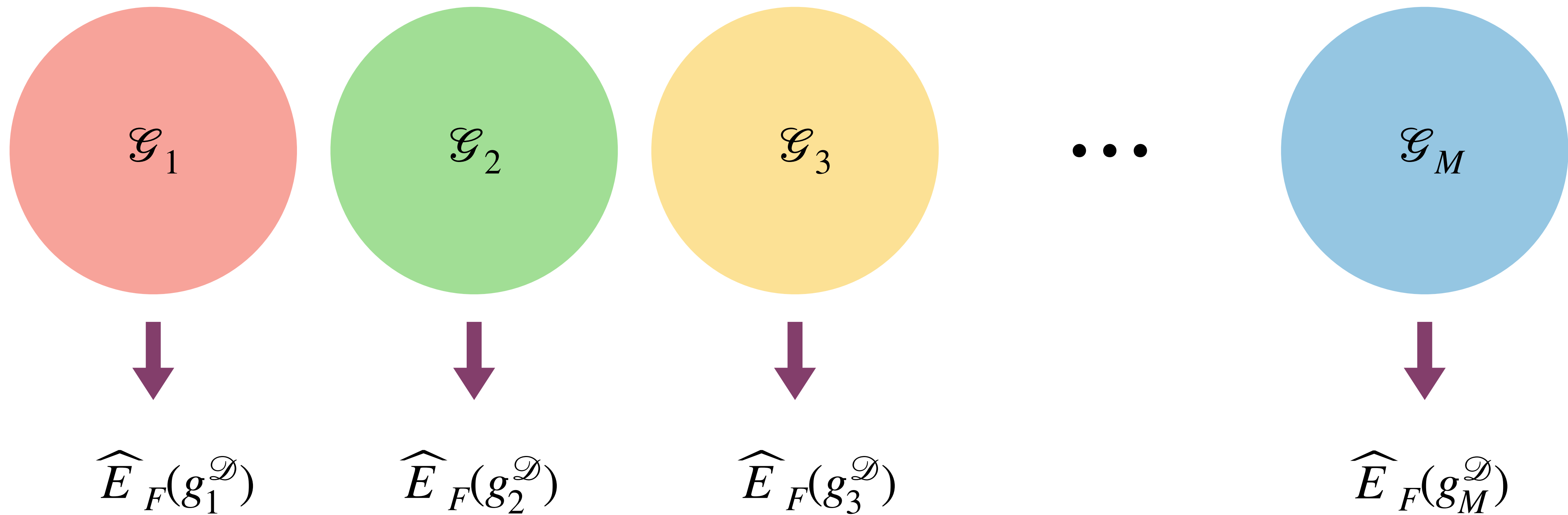
Neste caso, mesmo para  $g = g^{\mathcal{D}}$  temos que

$$\left. \begin{aligned} \mathbb{E}(\hat{E}_F(g)) &= E_F(g) \\ \text{Var}(\hat{E}_F(g)) &= \frac{\text{Var}[L(y, g(x))]}{K} = \frac{C}{K} \end{aligned} \right\} \longrightarrow E_F(g) = \hat{E}_F(g) \pm O_p\left(\frac{1}{\sqrt{K}}\right)$$

# Estimação do erro “fora da amostra”



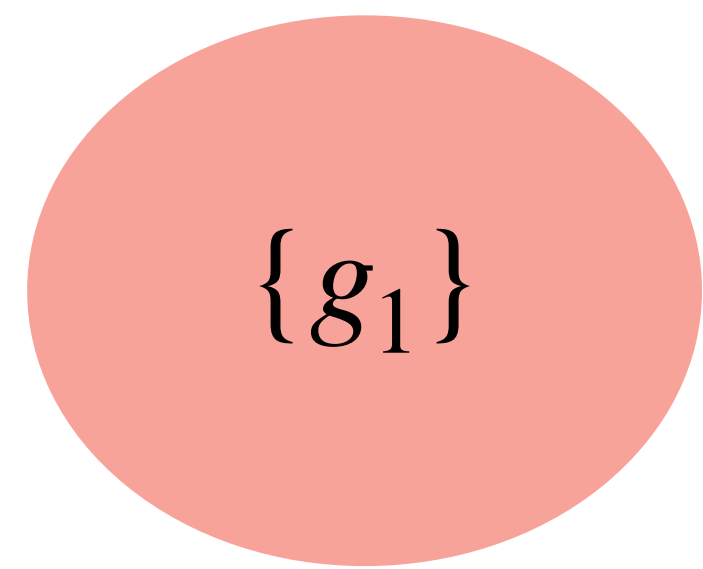
# Seleção de modelos



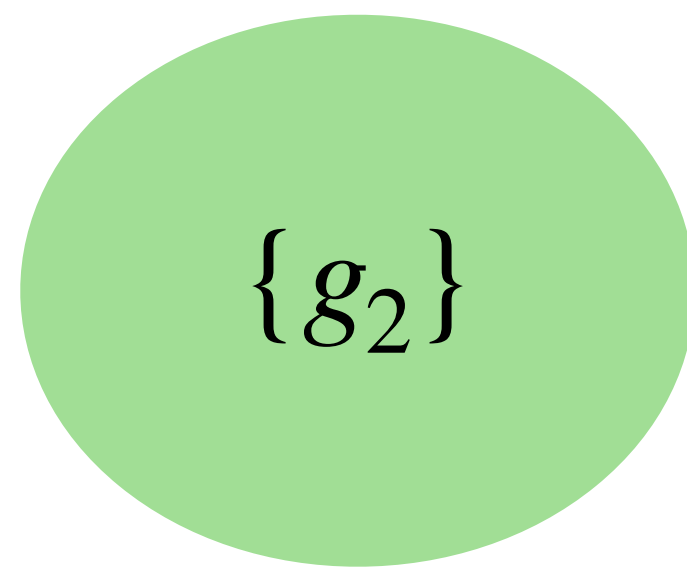
Escolhemos  $g_i^{\mathcal{D}}$  tal que  $\widehat{E}_F(g_i^{\mathcal{D}})$  é mínimo?

# Seleção de modelos

Escolhemos  $g_i^{\mathcal{D}}$  tal que  $\widehat{E}_F(g_i^{\mathcal{D}})$  é mínimo?



$$\widehat{E}_F(g_1)$$



$$\widehat{E}_F(g_2)$$

$$\widehat{E}_F(g_1), \widehat{E}_F(g_2) \sim \text{Uniforme}(0,1)$$

$$E_F(g_1) = E_F(g_2) = 0.5$$

$$g^{\mathcal{D}} = \arg \min(\widehat{E}_F(g_1), \widehat{E}_F(g_2))$$

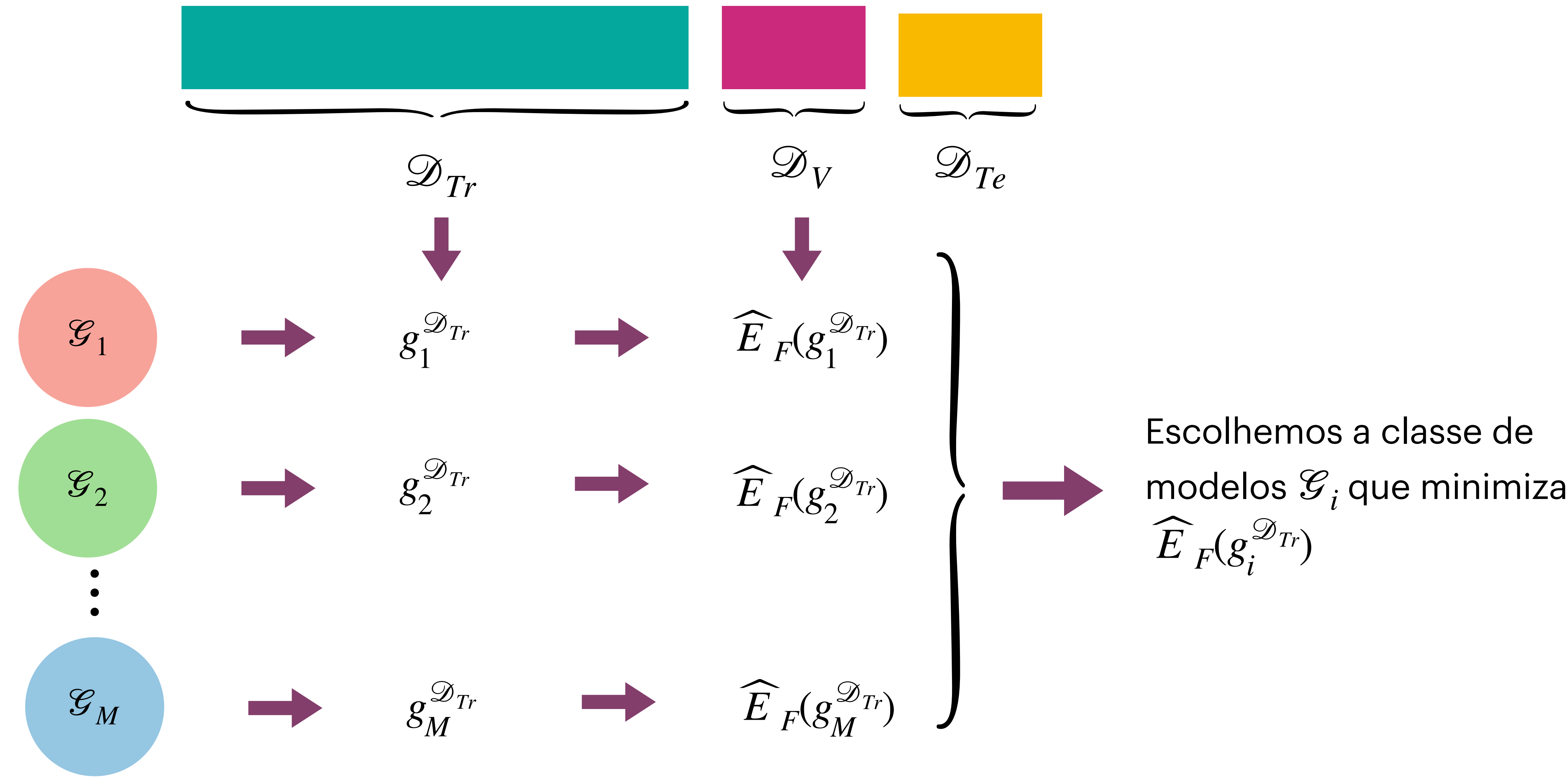
$$\mathbb{E}[\widehat{E}_F(g^{\mathcal{D}})] < 0.5 !$$

# Seleção de modelos

Quando um conjunto de dados de teste é utilizado para guiar a escolha do modelo, então ele passa a ser um conjunto de “validação”

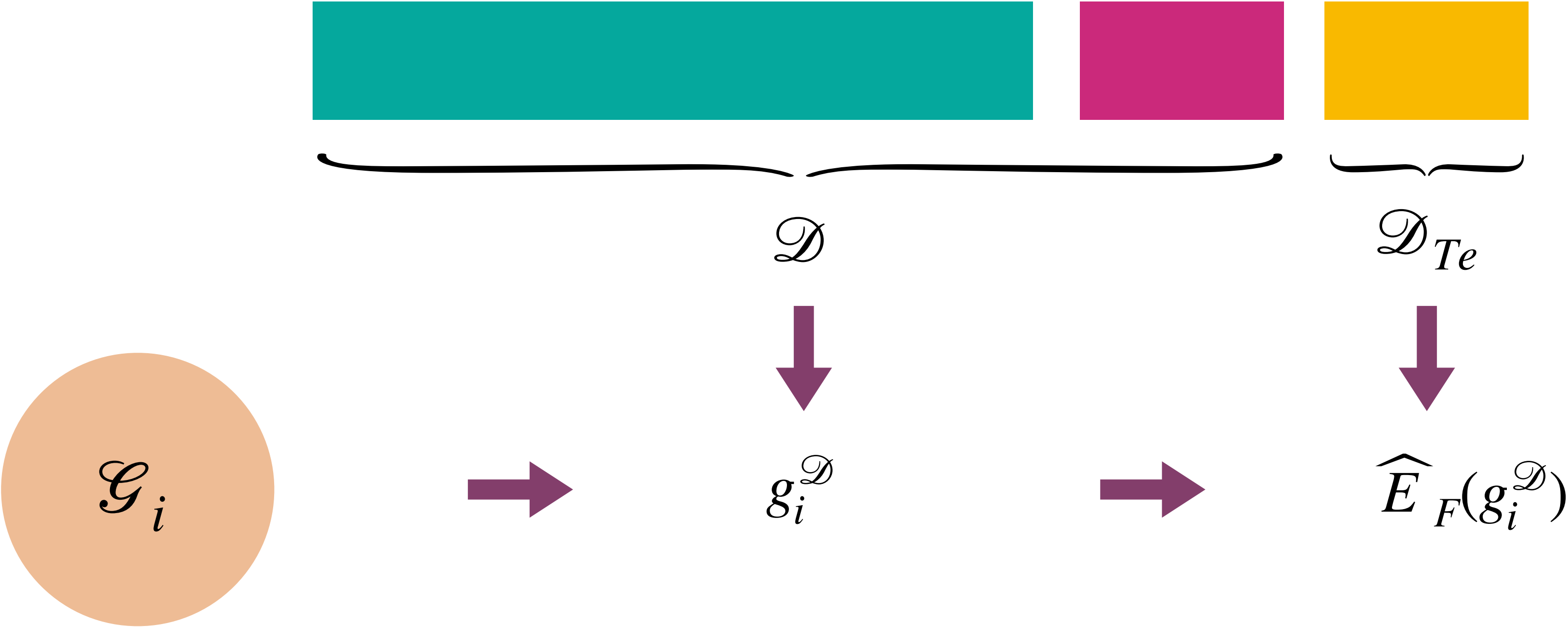


# Seleção de modelos



# Seleção de modelos

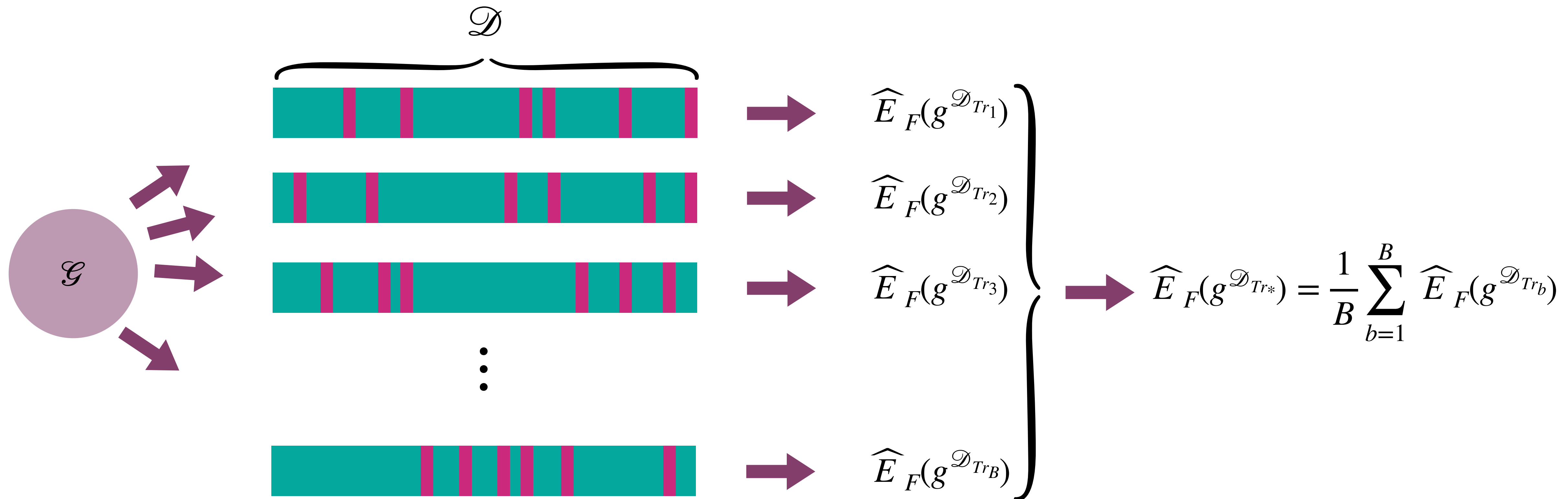
Se  $\mathcal{G}_i$  é a classe que minimiza  $\widehat{E}_F(g_i^{\mathcal{D}_{Tr}})$





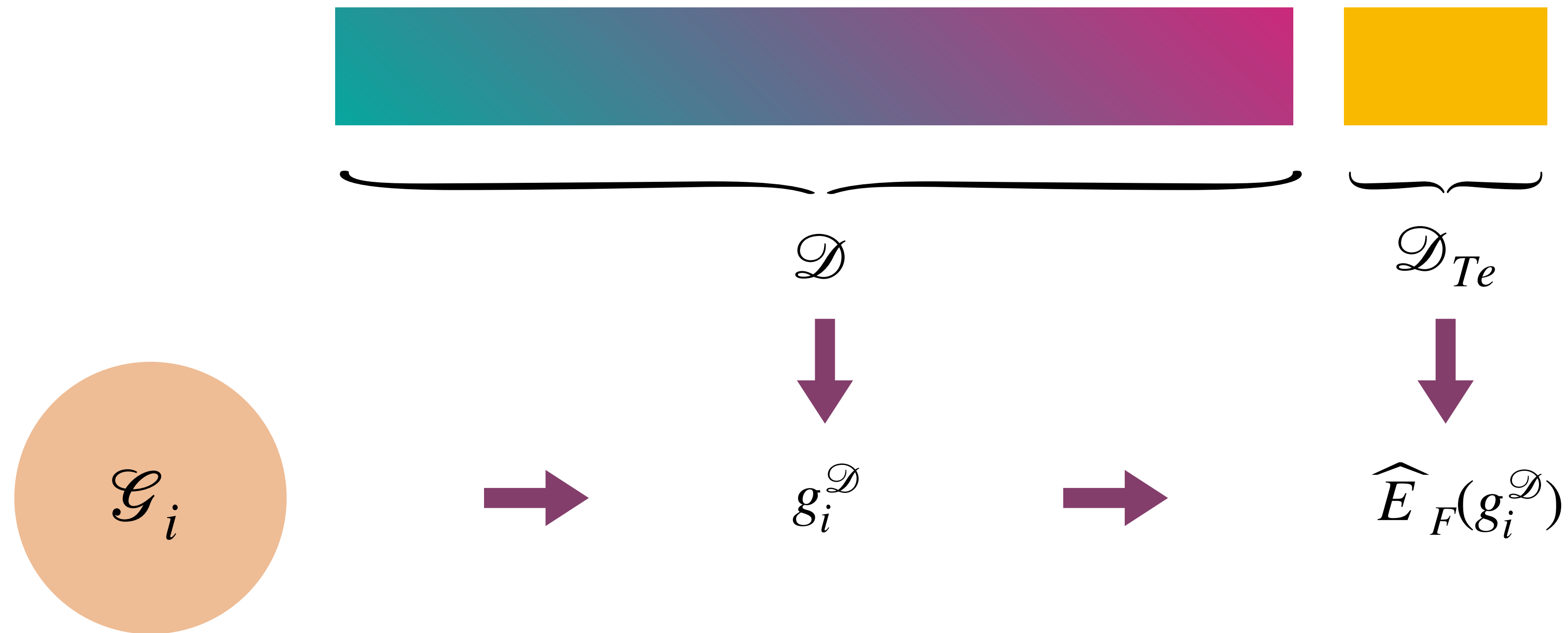
# Seleção de modelos por reamostragem (*bootstrap*)

O problema com o esquema treinamento/validação/teste é que temos pouca variabilidade na amostra de treinamento e validação (uma única amostra)



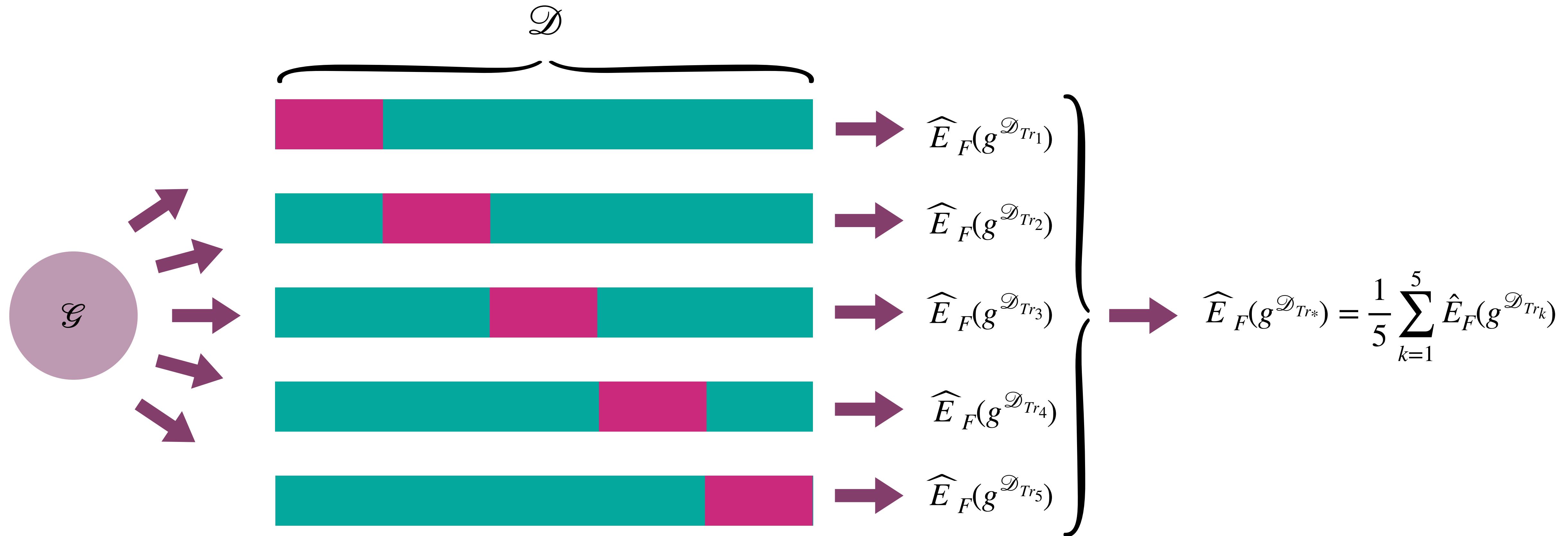
# Seleção de modelos por reamostragem (*bootstrap*)

Se  $\mathcal{G}_i$  é a classe que minimiza  $\widehat{E}_F(g_i^{\mathcal{D}_{Tr*}})$



# Seleção de modelos por validação cruzada

Divisão da amostra em  $k$  lotes do mesmo tamanho



Em geral  $k = 5$  ou  $k = 10$

# Seleção de modelos por validação cruzada

Se  $\mathcal{G}_i$  é a classe que minimiza  $\widehat{E}_F(g_i^{\mathcal{D}_{Tr*}})$

