Descomposición Sesgo - Varianza

M.Sc. William Caicedo Torres

Universidad Tecnológica de Bolívar

caicedo77@gmail.com

21 de octubre de 2016

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Material inspirado en el curso de Caltech “Learning from Data”, por Yaser S. Abu Mustafá.

Un Quid Pro Quo en el aprendizaje

- La cota VC:

  $P[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq 4 m_{\mathcal{H}}(2N) e^{-\frac{1}{8}\epsilon^2 N}$ (Vapnik-Chervonenkis)

- Entre más complejo el conjunto de hipótesis utilizado, menor E_{in}.
- Sin embargo, en virtud de la cota VC, menor complejidad significa una mejor oportunidad de generalizar: Menor E_{out}.
- Lo ideal sería tener la complejidad justa, es decir que $\mathcal{H} = \{f\}$.

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Cuantificando el Quid Pro Quo

- La aproximación del análisis VC descompone el error fuera del entrenamiento como: $E_{out} \leq E_{in} + \Omega$
- Hay otra forma de descomponer E_{out}: El análisis Sesgo-Varianza.
- Usando esta aproximación, E_{out} se compone de:
  1. Que tan bien aproxima $\mathcal{H}$ a f.
  2. Que tan fácil es escoger una buena hipótesis h dentro de $\mathcal{H}$.
- Lo interesante del análisis Sesgo-Varianza es que lo podemos utilizar fácilmente en algoritmos de **Regresión** usando el **Error Cuadrado Medio**.

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Descomposición Sesgo-Varianza

- Comenzemos con E_{out}:

$$E_{out}(g^{(\mathcal{D})}) = \mathbb{E}_x \left[(g^{(\mathcal{D})}(x) - f(x))^2 \right]$$

$$\mathbb{E}_{(\mathcal{D})} \left[E_{out}(g^{(\mathcal{D})}) \right] = \mathbb{E}_{(\mathcal{D})} \left[\mathbb{E}_x \left[(g^{(\mathcal{D})}(x) - f(x))^2 \right] \right]$$

$$\mathbb{E}_{(\mathcal{D})} \left[E_{out}(g^{(\mathcal{D})}) \right] = \mathbb{E}_x \left[\mathbb{E}_{(\mathcal{D})} \left[(g^{(\mathcal{D})}(x) - f(x))^2 \right] \right]$$

- Ahora, concentrémonos en:

$$\mathbb{E}_{(\mathcal{D})} \left[(g^{(\mathcal{D})}(x) - f(x))^2 \right]$$

La Hipótesis Promedio

- Para evaluar $\mathbb{E}_{(\mathcal{D})}\left[(g^{(\mathcal{D})}(x) - f(x))^2\right]$ debemos definir la hipótesis "promedio":

$$\overline{g}(x) = \mathbb{E}_{(\mathcal{D})}\left[g^{(\mathcal{D})}(x)\right]$$

- Qué es la hipótesis promedio? Imagine muchos datasets $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$
- Entonces:

$$\overline{g}(x) \approx \frac{1}{K}\sum_{k=1}^{K} g^{(\mathcal{D}_k)}(x)$$

- La hipótesis promedio es algo así como la "mejor" hipótesis que puedo encontrar en $\mathcal{H}$, a partir multiples datasets $\mathcal{D}_k$

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Usando $\overline{g}(x)$ en el análisis

$$\mathbb{E}_{(\mathcal{D})}\left[(g^{(\mathcal{D})}(x) - f(x))^2\right] = \mathbb{E}_{(\mathcal{D})}\left[(g^{(\mathcal{D})}(x) - \overline{g}(x) + \overline{g}(x) - f(x))^2\right]$$

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

$$= \mathbb{E}_{(\mathcal{D})}\left[(g^{(\mathcal{D})}(x) - \overline{g}(x))^2\right] + (\overline{g}(x) - f(x))^2$$

Notas: 1. El valor esperado se puede distribuir con respecto a la suma. 2. $\overline{g}(x) = \mathbb{E}_{(\mathcal{D})}\left[g^{(\mathcal{D})}(x)\right]$, por lo que la resta del primer factor del doble producto equivale a cero. 3. $\overline{g}(x) - f(x)$ es constante con respecto a $\mathcal{D}$, por lo que se puede factorizar fuera del valor esperado.

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Sesgo y Varianza

$$\mathbb{E}_{(\mathcal{D})}\left[(g^{(\mathcal{D})}(x) - f(x))^2\right] = \underbrace{\mathbb{E}_{(\mathcal{D})}\left[(g^{(\mathcal{D})}(x) - \overline{g}(x))^2\right]}_{\text{Varianza}} + \underbrace{(\overline{g}(x) - f(x))^2}_{\textit{Sesgo}}$$

Recordemos que,

$$\mathbb{E}_{(\mathcal{D})}\left[E_{out}(g^{(\mathcal{D})})\right] = \mathbb{E}_x\left[\mathbb{E}_{(\mathcal{D})}\left[(g^{(\mathcal{D})}(x) - f(x))^2\right]\right]$$

Por lo tanto,

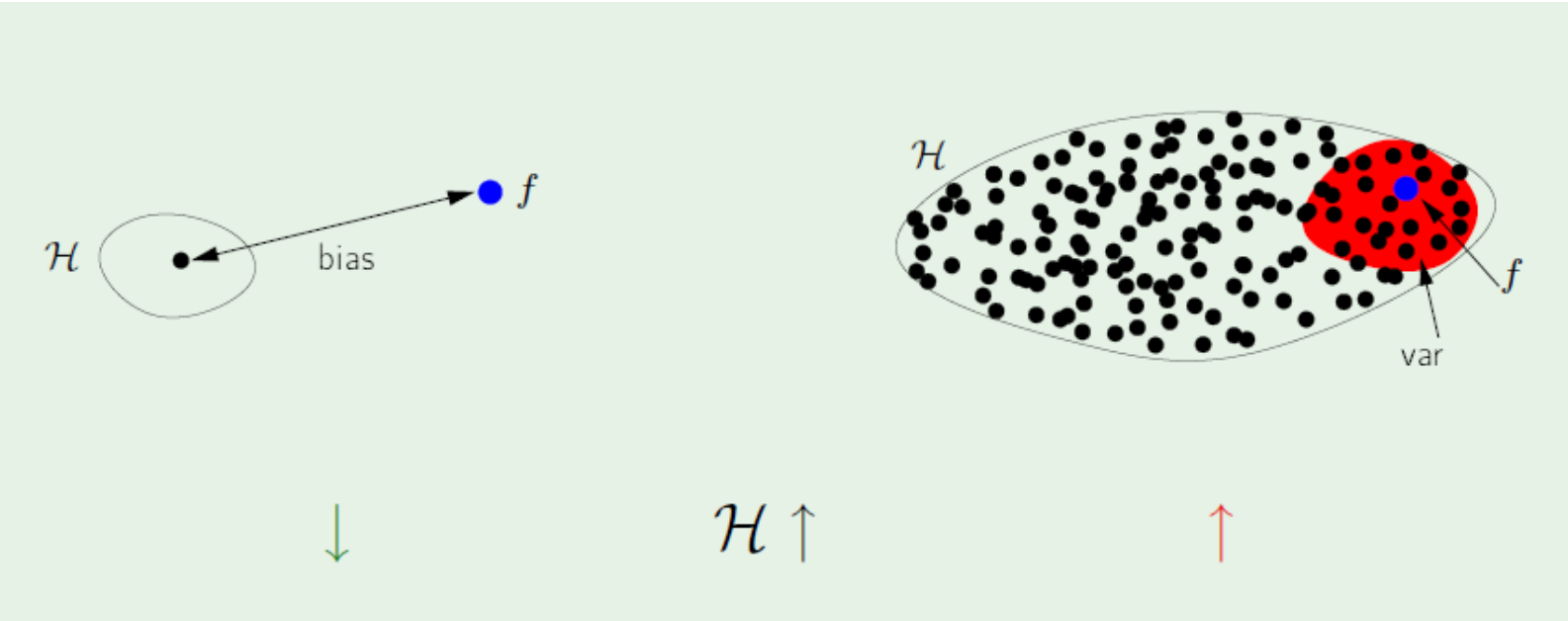
$$\mathbb{E}_{(\mathcal{D})}\left[E_{out}(g^{(\mathcal{D})})\right] = \mathbb{E}_x\left[Varianza(x) + Sesgo(x)\right]$$

$$= Varianza + Sesgo$$

El Quid Pro Quo según nuestra descomposición

$$sesgo = \mathbb{E}_x \left[(\overline{g}(x) - f(x))^2 \right]$$

$$varianza = \mathbb{E}_x \left[\mathbb{E}_{(\mathcal{D})} \left[(g^{(\mathcal{D})}(x) - \overline{g}(x))^2 \right] \right]$$



Ejemplo: Función Seno

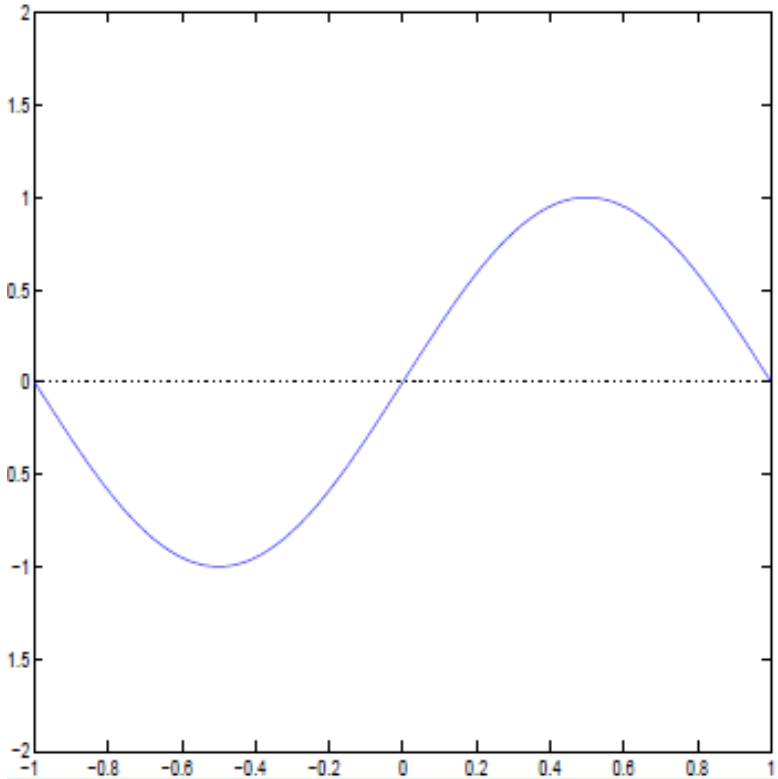
$f : [-1, 1] \rightarrow \mathbb{R} \quad f(x) = sen(\pi x)$

Solo disponemos de 2 ejemplos de entrenamiento Tenemos 2 modelos de de aprendizaje:

$$\mathcal{H}_0 : h(x) = b$$

$$\mathcal{H}_1 : h(x) = ax + b$$

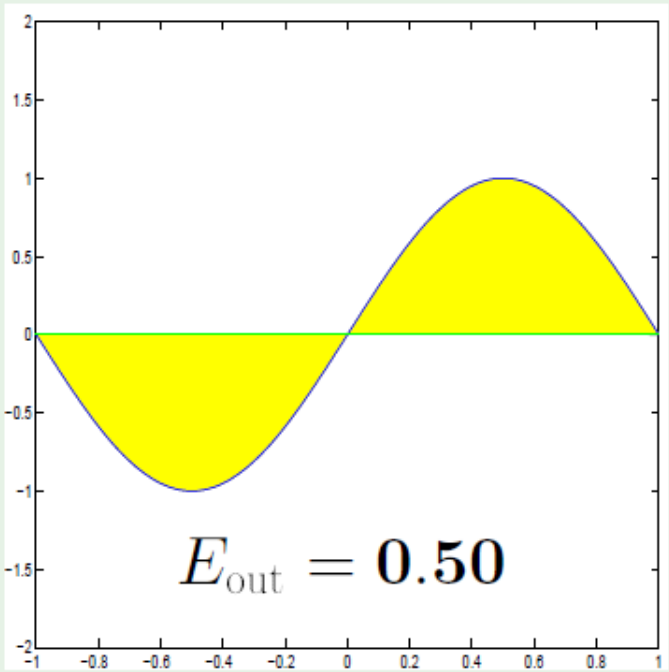
Cuál es mejor, $\mathcal{H}_0$ o $\mathcal{H}_1$?



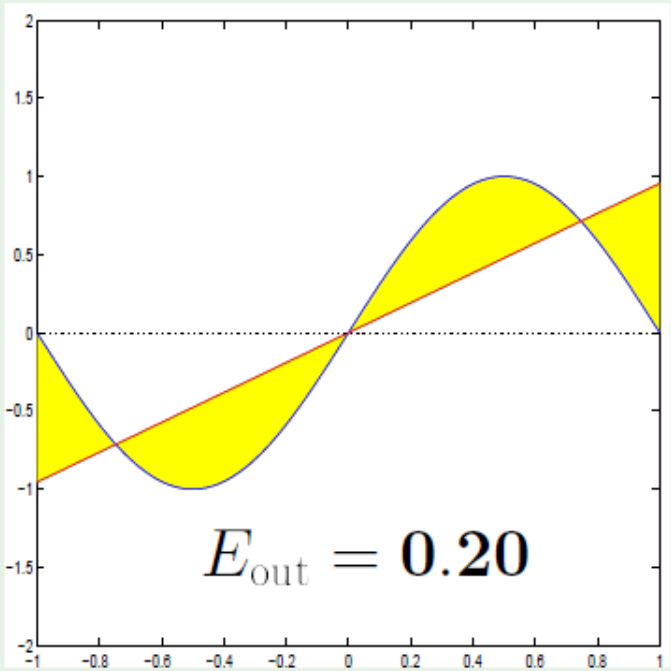
Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

En un mundo ideal...

$\mathcal{H}_0$



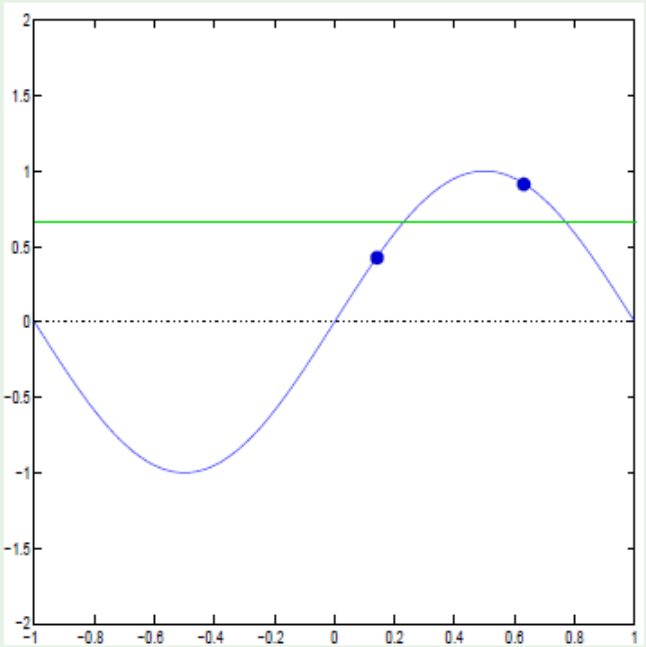
$\mathcal{H}_1$



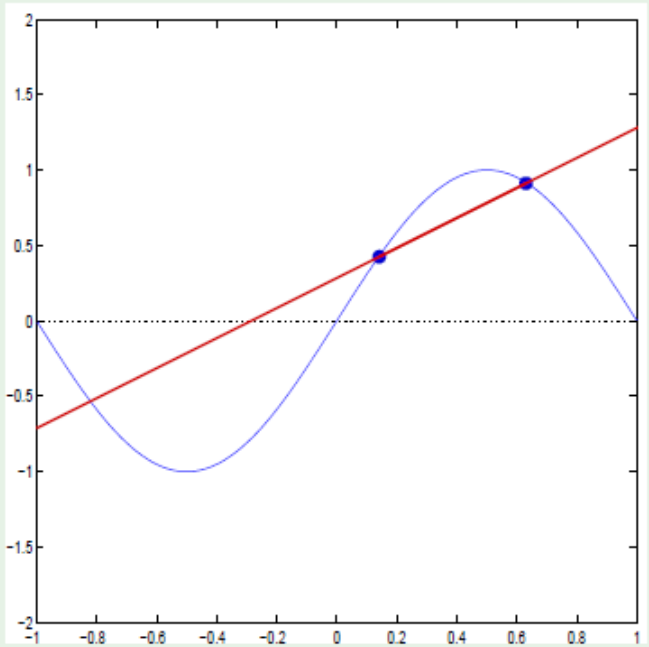
Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Pero solo tenemos 2 ejemplos de entrenamiento!

$\mathcal{H}_0$

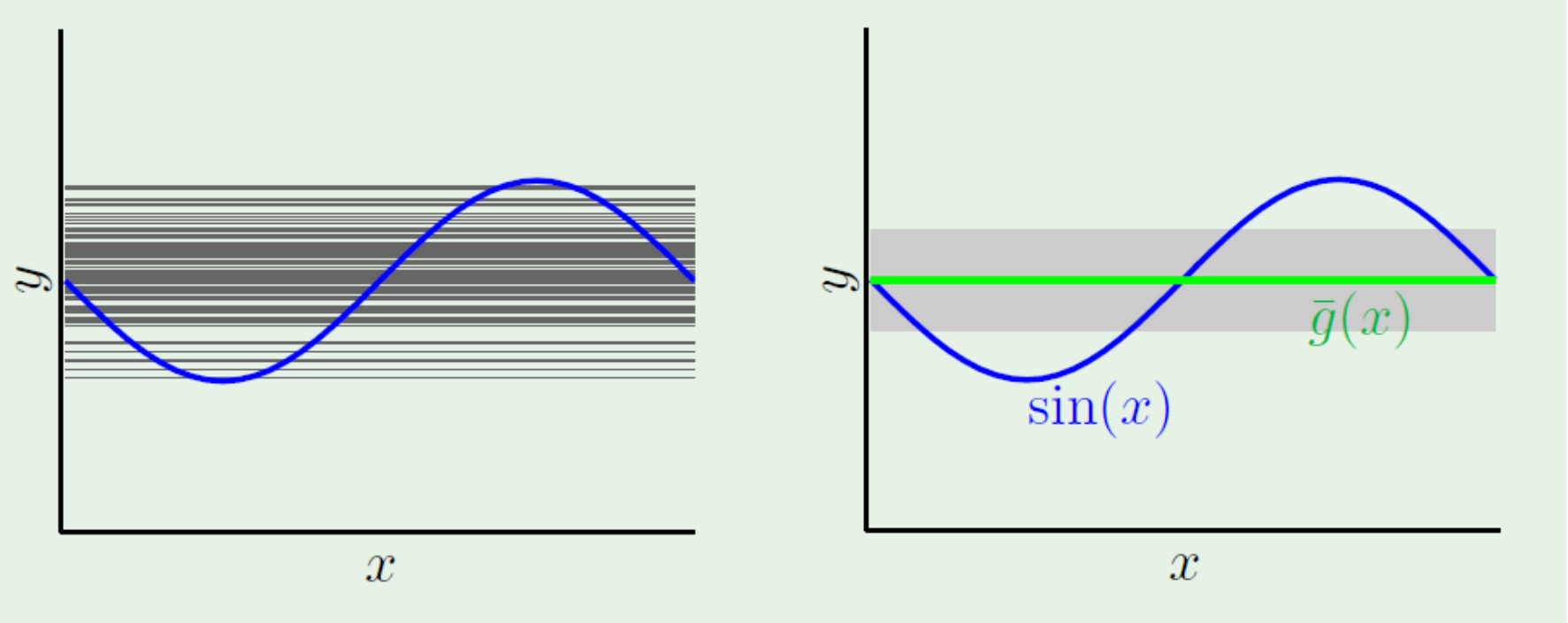


$\mathcal{H}_1$

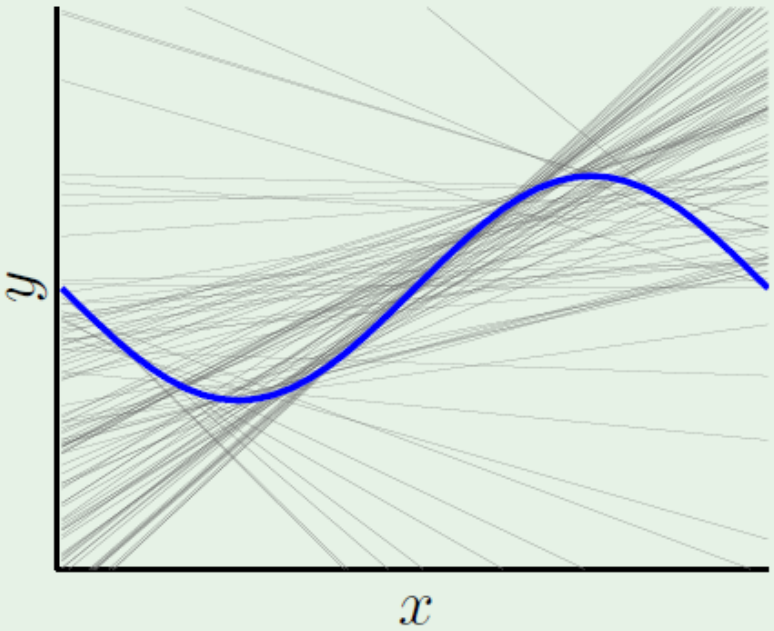


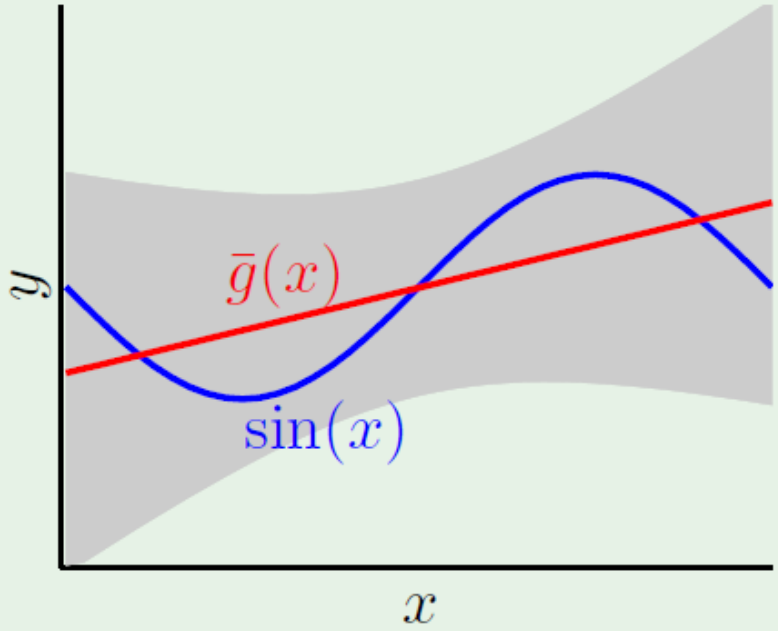
Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Sesgo y varianza para $\mathcal{H}_0$

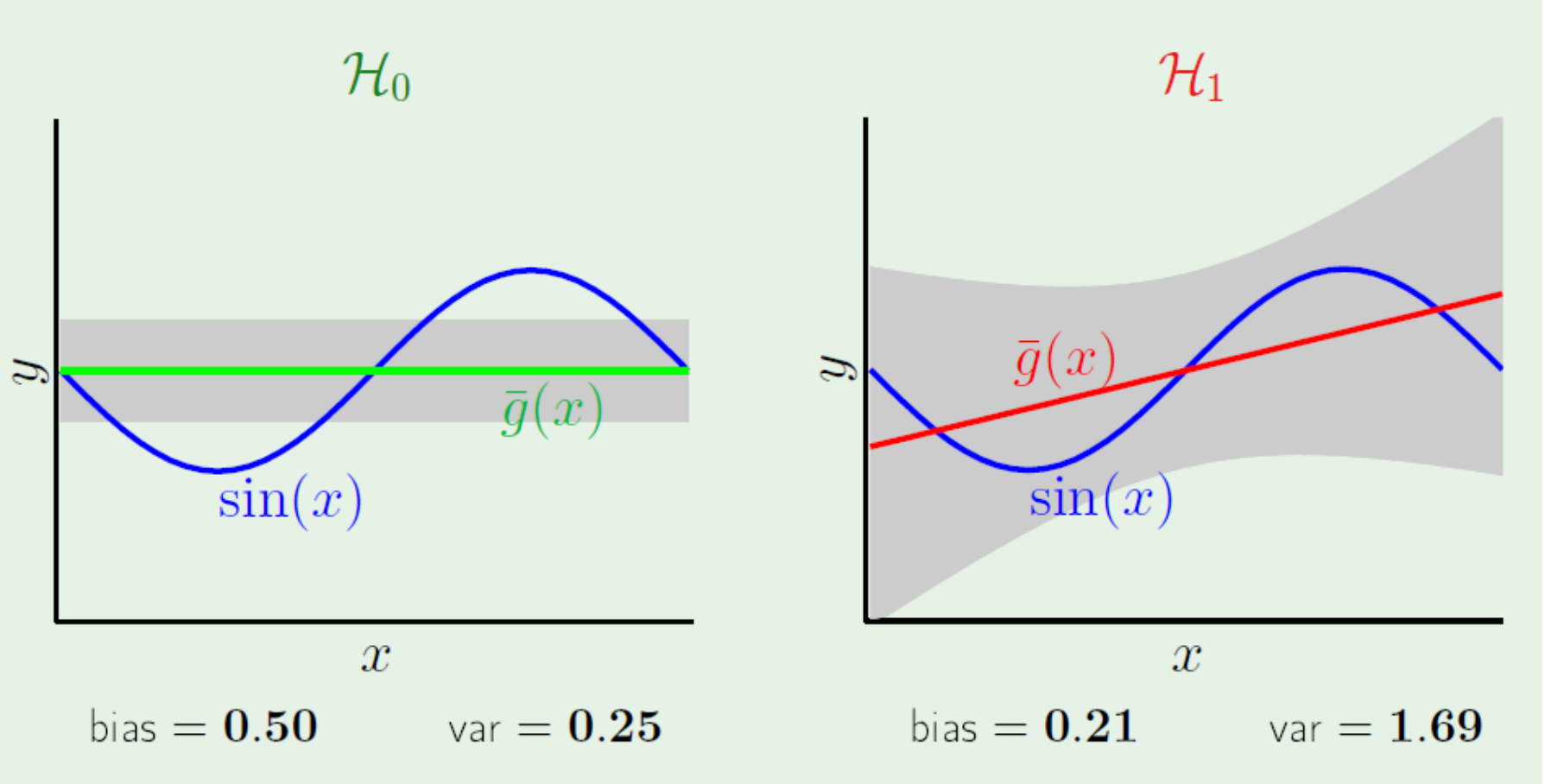


Sesgo y varianza para $\mathcal{H}_1$





Y el ganador es...



Lección de hoy

Decida la complejidad del modelo de aprendizaje a utilizar de acuerdo a los datos disponibles, no de acuerdo a la complejidad que ud supone tiene la función blanco!

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Muchas gracias!

Preguntas?

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS