Aprendizado Profundo 1

Overfitting, *Underfitting* e Regularização

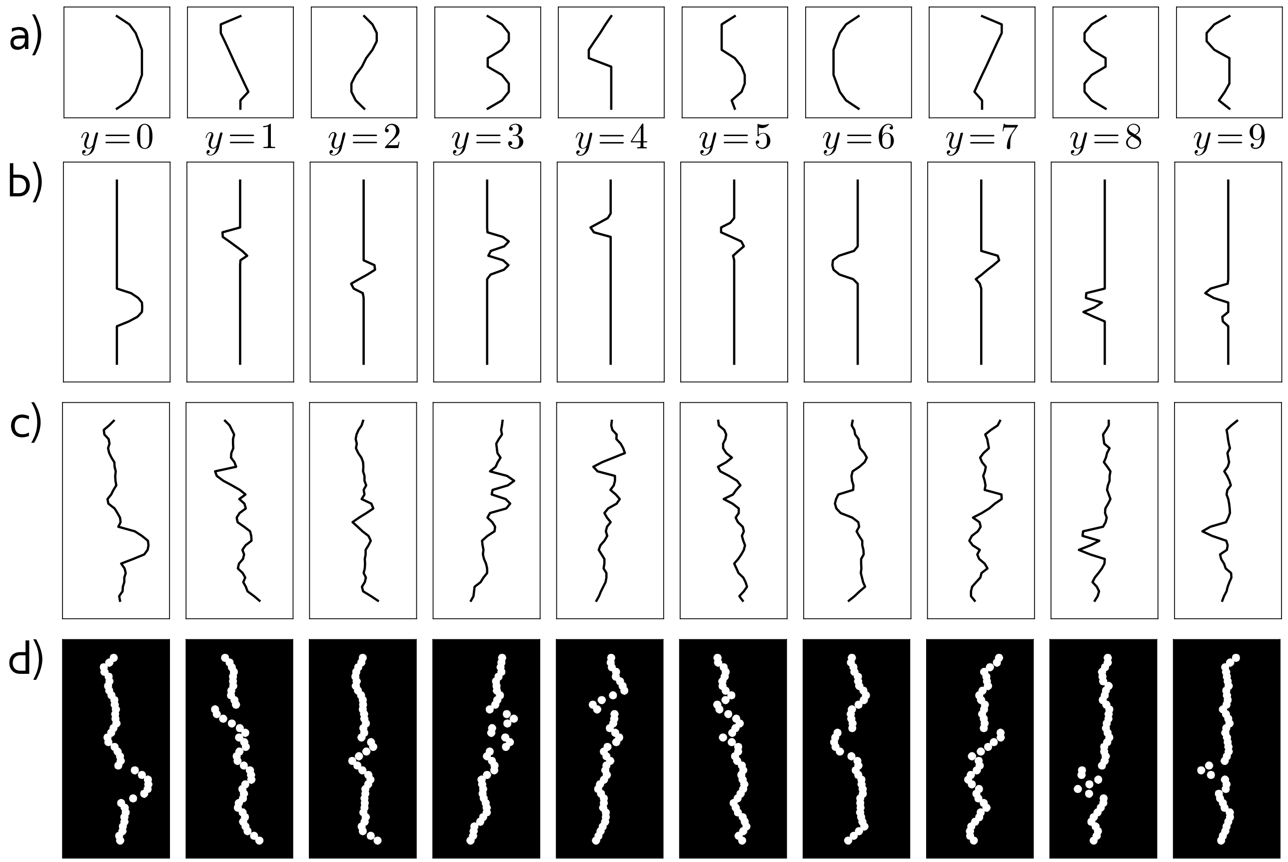
Professor: Lucas Silveira Kupssinskü

Agenda

- Fontes de Erro
  - *Noise*, *Bias* e *Variance*
- *Overfitting vs Underfitting*
  - *Bias Variance Tradeoff*
  - *Double Descent*
- Regularização
  - Métodos Explícitos
    - L1, L2
  - Métodos Empíricos
    - *Early Stopping*
    - *Dropout*
    - *Data Augmentation*
  - Métodos Implícitos

Conjunto de dados

- MNIST 1D
- a) Templates para as 10 classes
- b) Exemplos de treino
- c) Adição de ruído
- d) Amostragem em 40 pontos



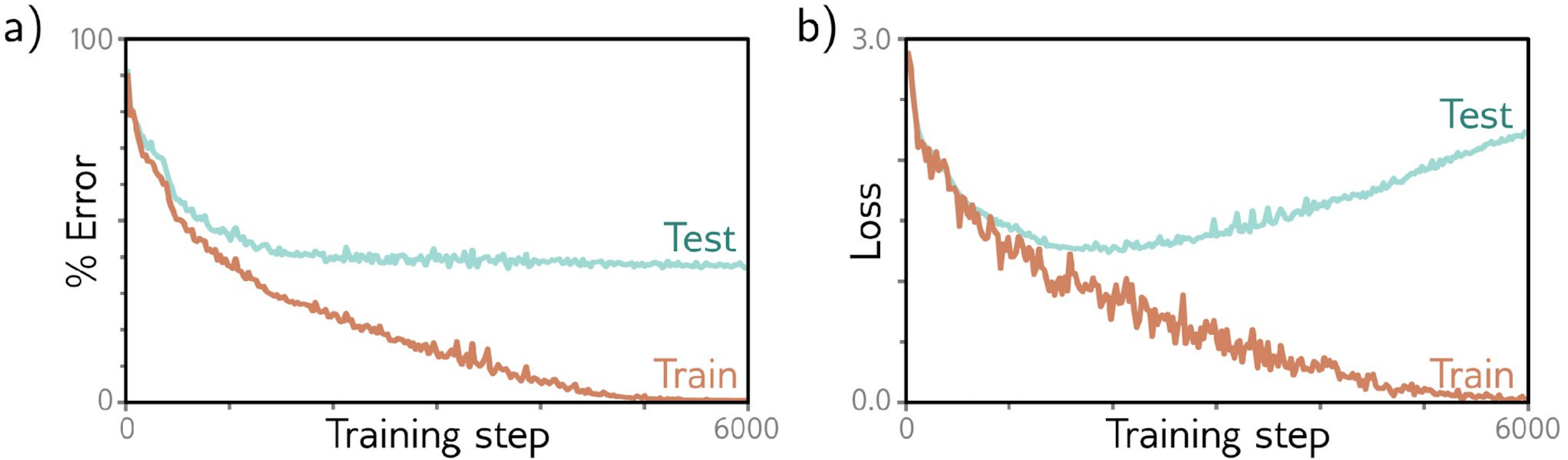
GREYDANUS, Sam. Scaling down deep learning. **arXiv preprint arXiv:2011.14439**, 2020.

https://github.com/greydanus/mnist1d

MLP

- 40 Entradas
- 10 Saídas (*Softmax*)
- 4000 instâncias de treino
- Duas camadas ocultas
  - 100 unidades em cada
- SGD com *batch size* 100, *learning rate* 0.1
- 6000 iterações
  - Quantas épocas?

Resultados do Treinamento

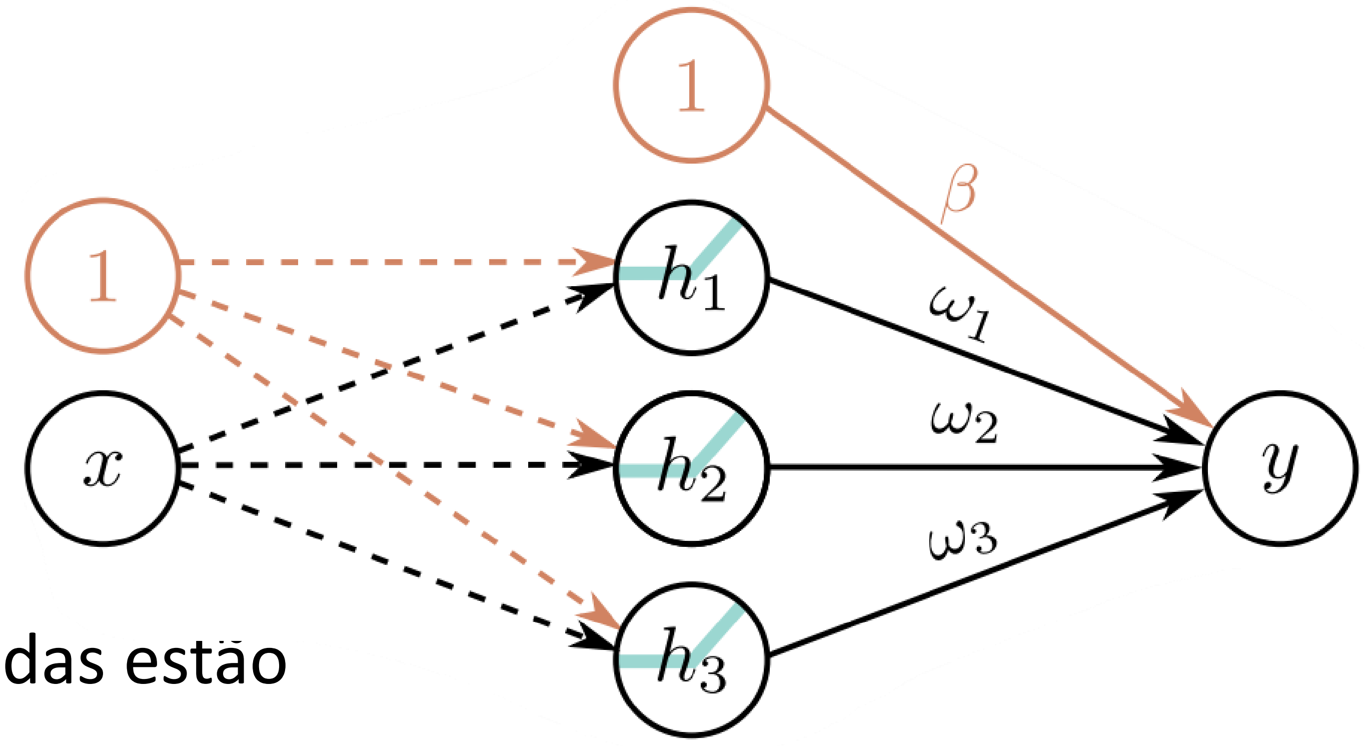


a) A taxa de erro em treinamento diminui até zero

b) A função de custo em teste passa a aumentar enquanto em treinamento ainda há diminuição

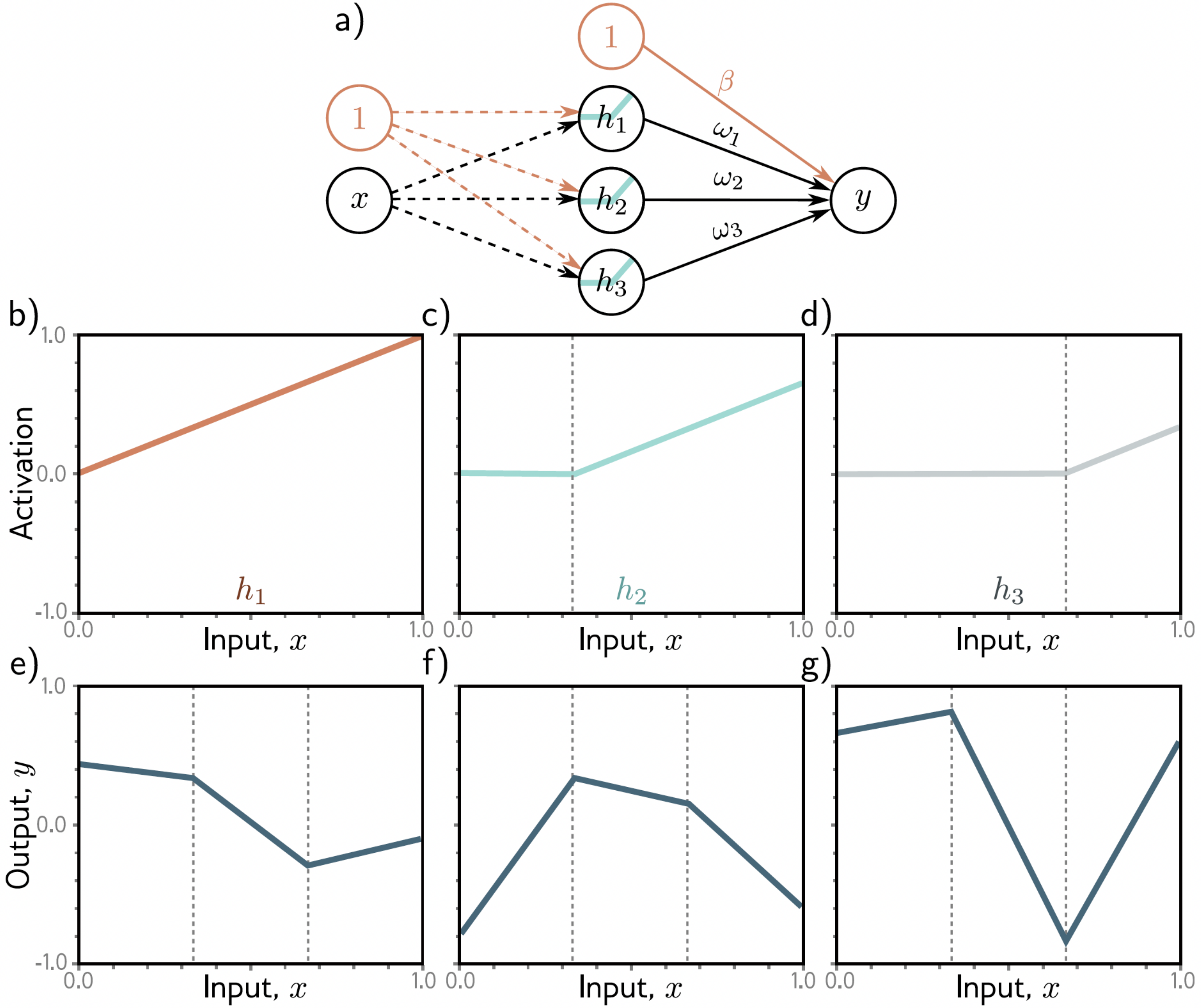
MLP

- 1 Entrada
- 1 Saída (*Linear*)
- Uma camada ocultas
  - 3 unidades em cada
  - ReLU
- Parâmetros nas linhas pontilhadas estão congelados



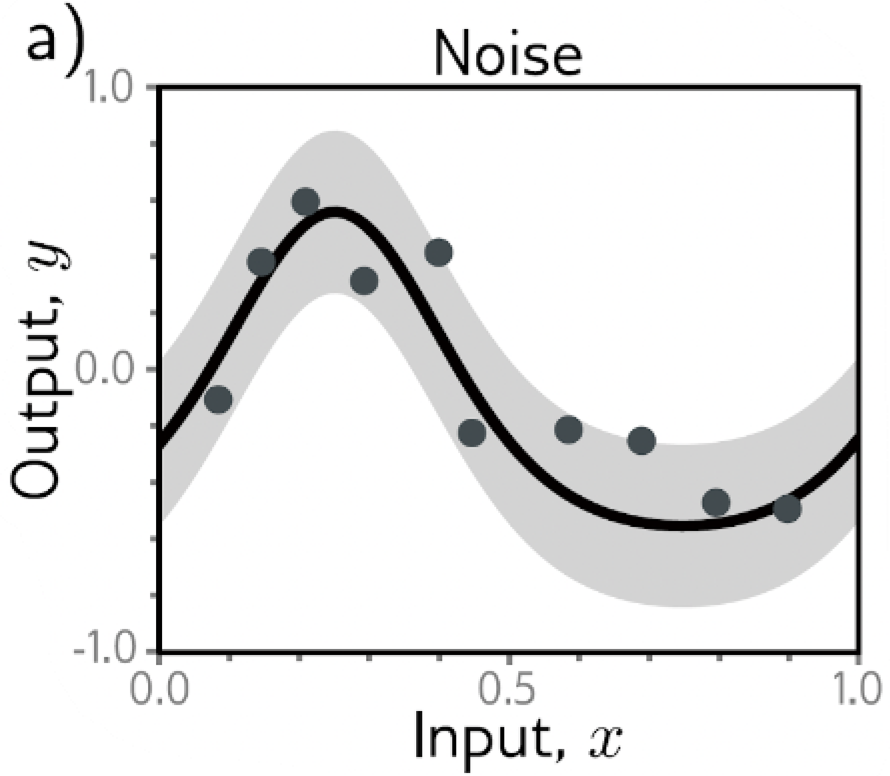
MLP

- Repare em e), f) e g) diferentes parametrizações geram funções lineares diferentes



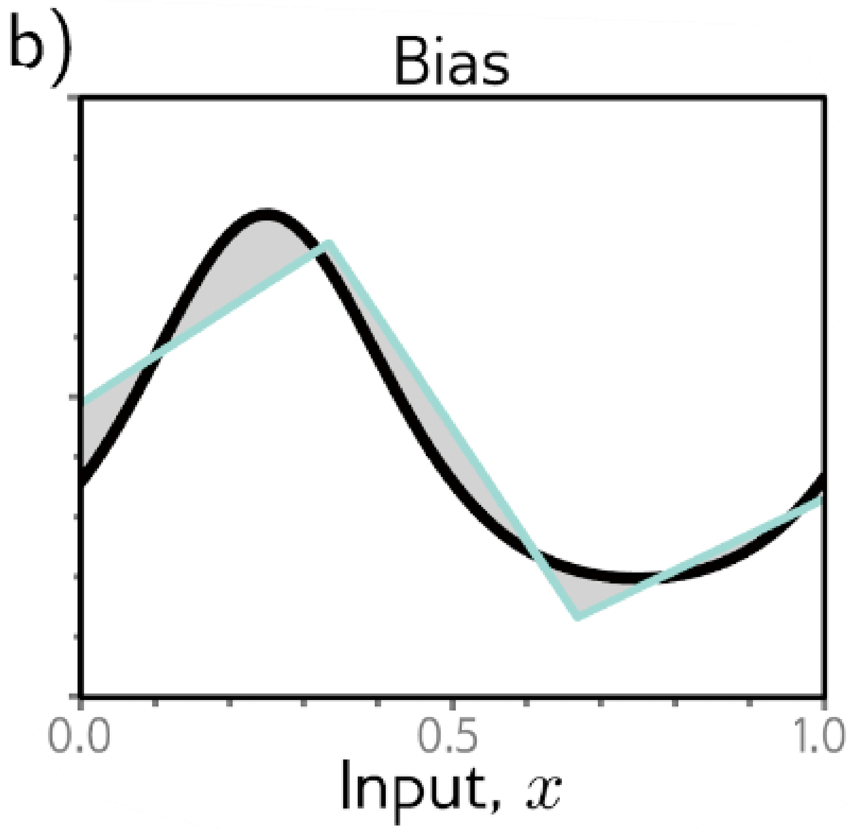
Fontes de Erro

- Ruído
  - Por mais que consigamos modelar a função geradora dos dados (sinusoide em preto), ainda teremos um erro diferente de 0
  - Fatores estocásticos ou ausência de uma variável



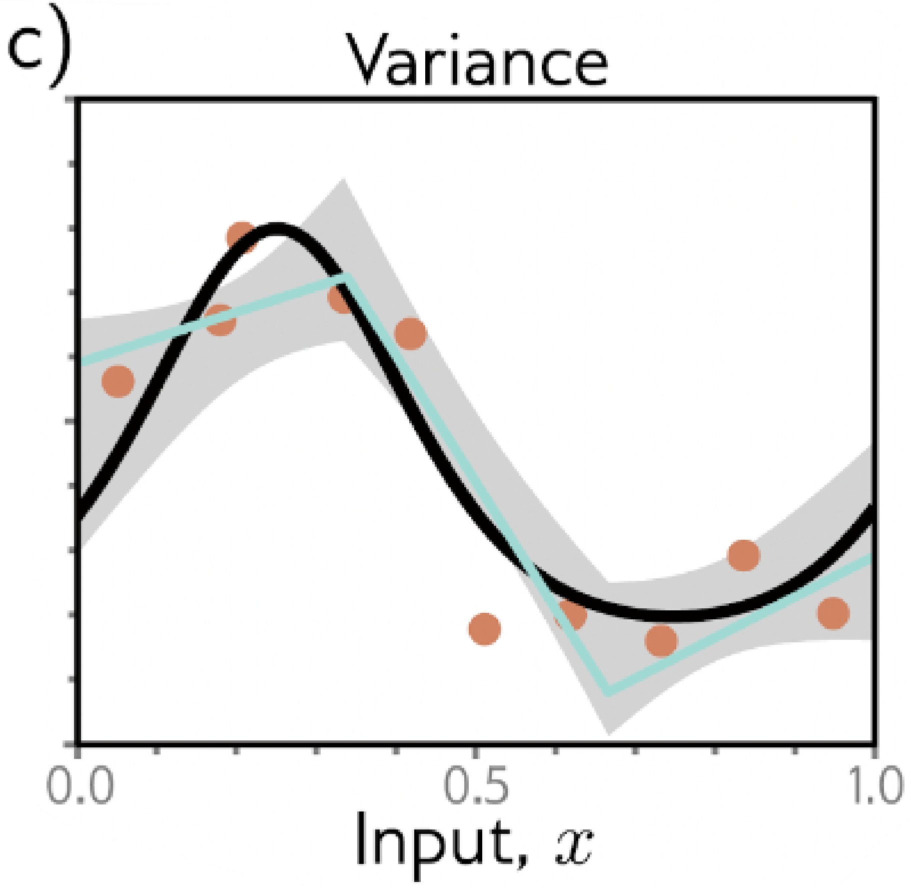
Fontes de Erro

- *Bias*
  - Mesmo com os melhores parâmetros nosso MLP (cyan) não consegue modelar perfeitamente a função em preto
  - Todo modelo precisa de viés indutivo para funcionar

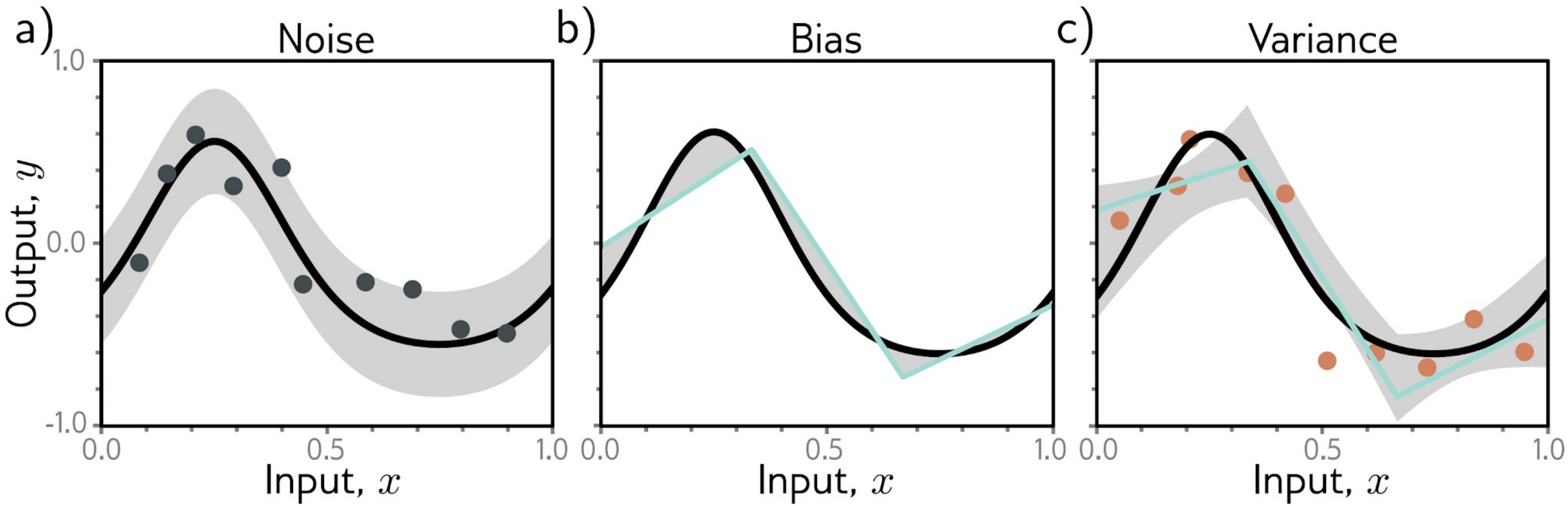


Fontes de Erro

- *Variance*
  - Quando treinamos o modelo em um número limitado de dados, não recuperamos necessariamente o melhor ajuste a função geradora



Fontes de Erro



Fontes de Erro

- Vamos tornar precisa a noção da fonte de erro
- Considere um problema de regressão de uma variável
- Suponha que:
  - Para valores distintos de x temos valores esperados diferentes, ou seja $\mathbb{E}[\Pr(y|x)] = \mu[x]$
  - Existe um ruído fixo $\sigma^2 = \mathbb{E}[(\mu[x]\ - \mathrm{y}[\mathrm{x}])^2]$
  - A predição do modelo é denotada por $f[x, \theta]$
  - Vamos considerar o erro quadrático como função de custo $L[x] = (f[x, \theta] - \mathrm{y}[\mathrm{x}])^2$

Fontes de Erro

- Suponha que:
  - Para valores distintos de x temos valores esperados diferentes, ou seja $\mathbb{E}[\mathrm{Pr}(y|x)] = \mu[x]$
  - Existe um ruído fixo $\sigma^2 = \mathbb{E}[(\mu[x] \; - \mathrm{y}[\mathrm{x}])^2]$
  - A predição do modelo é denotada por $f[x, \theta]$
  - Vamos considerar o erro quadrático como função de custo $L[x] = (f[x, \theta] - \mathrm{y}[\mathrm{x}])^2$

$$L[x] = (f[x, \theta] - \mathrm{y}[\mathrm{x}])^2$$

Fontes de Erro

- Suponha que:
  - Para valores distintos de x temos valores esperados diferentes, ou seja $\mathbb{E}[\mathrm{Pr}(y|x)] = \mu[x]$
  - Existe um ruído fixo $\sigma^2 = \mathbb{E}[(\mu[x] \ - \mathrm{y}[\mathrm{x}])^2]$
  - A predição do modelo é denotada por $f[x, \theta]$
  - Vamos considerar o erro quadrático como função de custo $L[x] = (f[x, \theta] - \mathrm{y}[\mathrm{x}])^2$

$$
\begin{aligned}
L[x] &= (f[x, \theta] - \mathrm{y}[\mathrm{x}])^2 \\
&= \big((f[x, \theta] - \mu[x]) + (\mu[x] - \mathrm{y}[\mathrm{x}])\big)^2
\end{aligned}
$$

Fontes de Erro

- Suponha que:
  - Para valores distintos de x temos valores esperados diferentes, ou seja $\mathbb{E}[\Pr(y|x)] = \mu[x]$
  - Existe um ruído fixo $\sigma^2 = \mathbb{E}[(\mu[x] - \mathrm{y}[\mathrm{x}])^2]$
  - A predição do modelo é denotada por $f[x, \theta]$
  - Vamos considerar o erro quadrático como função de custo $L[x] = (f[x, \theta] - \mathrm{y}[\mathrm{x}])^2$

$$
\begin{aligned}
L[x] &= (f[x, \theta] - \mathrm{y}[\mathrm{x}])^2 \\
&= \big((f[x, \theta] - \mu[x]) + (\mu[x] - \mathrm{y}[\mathrm{x}])\big)^2 \\
&= \big((f[x, \theta] - \mu[x])^2 + 2(f[x, \theta] - \mu[x])(\mu[x] - \mathrm{y}[\mathrm{x}]) + (\mu[x] - \mathrm{y}[\mathrm{x}])^2\big)
\end{aligned}
$$

Fontes de Erro

- Agora vamos verificar qual é o valor esperado da função de custo em relação a um valor y

$$
\begin{aligned}
L[x] &= \left((f[x,\theta]-\mu[x])^2 + 2(f[x,\theta]-\mu[x])(\mu[x]-\mathrm{y}[\mathrm{x}]) + (\mu[x]-\mathrm{y}[\mathrm{x}])^2\right)\\
\mathbb{E}_y\left[L[x]\right] &= \mathbb{E}_y\left[\left((f[x,\theta]-\mu[x])^2 + 2(f[x,\theta]-\mu[x])(\mu[x]-\mathrm{y}[\mathrm{x}]) + (\mu[x]-\mathrm{y}[\mathrm{x}])^2\right)\right]\\
\mathbb{E}_y\left[L[x]\right] &= \boxed{\mathbb{E}_y[(f[x,\theta]-\mu[x])^2]} + \boxed{\mathbb{E}_y[2(f[x,\theta]-\mu[x])}(\boxed{\mu[x]}-\mathrm{y}[\mathrm{x}])] + \mathbb{E}_y[(\boxed{\mu[x]}-\mathrm{y}[\mathrm{x}])^2]
\end{aligned}
$$

Não dependem de y

Fontes de Erro

- Agora vamos verificar qual é o valor esperado da função de custo em relação a um valor y

$$
\begin{aligned}
L[x] &= \left((f[x,\theta]-\mu[x])^2 + 2(f[x,\theta]-\mu[x])(\mu[x]-\mathrm{y}[\mathrm{x}]) + (\mu[x]-\mathrm{y}[\mathrm{x}])^2\right) \\
\mathbb{E}_y\big[L[x]\big] &= \mathbb{E}_y\Big[\left((f[x,\theta]-\mu[x])^2 + 2(f[x,\theta]-\mu[x])(\mu[x]-\mathrm{y}[\mathrm{x}]) + (\mu[x]-\mathrm{y}[\mathrm{x}])^2\right)\Big] \\
\mathbb{E}_y\big[L[x]\big] &= \mathbb{E}_y[(f[x,\theta]-\mu[x])^2] + \mathbb{E}_y[2(f[x,\theta]-\mu[x])(\mu[x]-\mathrm{y}[\mathrm{x}])] + \mathbb{E}_y[(\mu[x]-\mathrm{y}[\mathrm{x}])^2] \\
\mathbb{E}_y\big[L[x]\big] &= (f[x,\theta]-\mu[x])^2 + 2(f[x,\theta]-\mu[x])\big(\mu[x] - \underset{\mu[x]}{\cancel{\mathbb{E}_y\big[\mathrm{y}[\mathrm{x}]\big]}}\big) + \mathbb{E}_y[(\mu[x]-\mathrm{y}[\mathrm{x}])^2]
\end{aligned}
$$

Fontes de Erro

- Agora vamos verificar qual é o valor esperado da função de custo em relação a um valor y

$$
\begin{aligned}
L[x] &= \left((f[x,\theta]-\mu[x])^2 + 2(f[x,\theta]-\mu[x])(\mu[x]-\mathrm{y}[\mathrm{x}]) + (\mu[x]-\mathrm{y}[\mathrm{x}])^2\right) \\
\mathbb{E}_y\big[L[x]\big] &= \mathbb{E}_y\left[\left((f[x,\theta]-\mu[x])^2 + 2(f[x,\theta]-\mu[x])(\mu[x]-\mathrm{y}[\mathrm{x}]) + (\mu[x]-\mathrm{y}[\mathrm{x}])^2\right)\right] \\
\mathbb{E}_y\big[L[x]\big] &= \mathbb{E}_y[(f[x,\theta]-\mu[x])^2] + \mathbb{E}_y[2(f[x,\theta]-\mu[x])(\mu[x]-\mathrm{y}[\mathrm{x}])] + \mathbb{E}_y[(\mu[x]-\mathrm{y}[\mathrm{x}])^2] \\
\mathbb{E}_y\big[L[x]\big] &= (f[x,\theta]-\mu[x])^2 + 2(f[x,\theta]-\mu[x])\big(\mu[x]-\mathbb{E}_y\big[\mathrm{y}[\mathrm{x}]\big]\big) + \mathbb{E}_y[(\mu[x]-\mathrm{y}[\mathrm{x}])^2] \\
\mathbb{E}_y\big[L[x]\big] &= (f[x,\theta]-\mu[x])^2 + 2(f[x,\theta]-\mu[x]){\color{red}0} + \cancel{\mathbb{E}_y[(\mu[x]-\mathrm{y}[\mathrm{x}])^2]}\ {\color{red}\sigma^2}
\end{aligned}
$$

Fontes de Erro

- Agora vamos verificar qual é o valor esperado da função de custo em relação a um valor y

$$
\begin{aligned}
L[x] &= \left((f[x,\theta]-\mu[x])^2 + 2(f[x,\theta]-\mu[x])(\mu[x]-\mathrm{y}[\mathrm{x}]) + (\mu[x]-\mathrm{y}[\mathrm{x}])^2\right)\\
\mathbb{E}_y\big[L[x]\big] &= \mathbb{E}_y\left[\left((f[x,\theta]-\mu[x])^2 + 2(f[x,\theta]-\mu[x])(\mu[x]-\mathrm{y}[\mathrm{x}]) + (\mu[x]-\mathrm{y}[\mathrm{x}])^2\right)\right]\\
\mathbb{E}_y\big[L[x]\big] &= \mathbb{E}_y[(f[x,\theta]-\mu[x])^2] + \mathbb{E}_y[2(f[x,\theta]-\mu[x])(\mu[x]-\mathrm{y}[\mathrm{x}])] + \mathbb{E}_y[(\mu[x]-\mathrm{y}[\mathrm{x}])^2]\\
\mathbb{E}_y\big[L[x]\big] &= (f[x,\theta]-\mu[x])^2 + 2(f[x,\theta]-\mu[x])\left(\mu[x]-\mathbb{E}_y\big[\mathrm{y}[\mathrm{x}]\big]\right) + \mathbb{E}_y[(\mu[x]-\mathrm{y}[\mathrm{x}])^2]\\
\mathbb{E}_y\big[L[x]\big] &= (f[x,\theta]-\mu[x])^2 + 2(f[x,\theta]-\mu[x])0 + \mathbb{E}_y[(\mu[x]-\mathrm{y}[\mathrm{x}])^2]\\
\mathbb{E}_y\big[L[x]\big] &= \boxed{(f[x,\theta]-\mu[x])^2} + \boxed{\sigma^2}
\end{aligned}
$$

Este termo se refere a nossas predições. Aqui podemos trabalhar mais

Este termo se refere ao ruído. Não temos como diminuir essa componente

Fontes de Erro

- Agora vamos verificar qual é o valor esperado da função de custo em relação a um valor y
  - Vamos definir um modelo médio em relação a todos os possíveis *datasets* $f_\mu[x] = \mathbb{E}_D\left[f\left[x, \theta[D]\right]\right]$

$$\mathbb{E}_y\left[L[x]\right] = \boxed{(f[x,\theta] - \mu[x])^2} + \sigma^2$$

$$(f[x,\theta] - \mu[x])^2 = \left(\left(f[x,\theta] - f_\mu[x]\right) + \left(f_\mu[x] - \mu[x]\right)\right)^2$$

$$= \left(f[x,\theta] - f_\mu[x]\right)^2 + 2\left(f[x,\theta] - f_\mu[x]\right)\left(f_\mu[x] - \mu[x]\right) + \left(f_\mu[x] - \mu[x]\right)^2$$

Fontes de Erro

- Agora vamos verificar qual é o valor esperado da função de custo em relação a um valor y
  - Vamos definir um modelo médio em relação a todos os possíveis *datasets* $f_\mu[x] = \mathbb{E}_D\left[f[x,\theta[D]]\right]$

$$\mathbb{E}_y\left[L[x]\right] = \boxed{(f[x,\theta] - \mu[x])^2} + \sigma^2$$

$$(f[x,\theta] - \mu[x])^2 = \left(\left(f[x,\theta] - f_\mu[x]\right) + \left(f_\mu[x] - \mu[x]\right)\right)^2$$

$$= \left(f[x,\theta] - f_\mu[x]\right)^2 + 2\left(f[x,\theta] - f_\mu[x]\right)\left(f_\mu[x] - \mu[x]\right) + \left(f_\mu[x] - \mu[x]\right)^2$$

$$\mathbb{E}_D[(f[x,\theta] - \mu[x])]^2 = \mathbb{E}_D\left[\left(f[x,\theta] - f_\mu[x]\right)^2 + 2\left(f[x,\theta] - f_\mu[x]\right)\left(f_\mu[x] - \mu[x]\right) + \left(f_\mu[x] - \mu[x]\right)^2\right]$$

$$= \mathbb{E}_D\left[\left(f[x,\theta] - f_\mu[x]\right)^2\right] + \mathbb{E}_D\left[2\left(f[x,\theta] - f_\mu[x]\right)\left(f_\mu[x] - \mu[x]\right)\right] + \mathbb{E}_D\left[\left(f_\mu[x] - \mu[x]\right)^2\right]$$

$$\mathbb{E}_D[(f[x,\theta] - \mu[x])]^2 = \mathbb{E}_D\left[\left(f[x,\theta] - f_\mu[x]\right)^2\right] + \left(f_\mu[x] - \mu[x]\right)^2$$

Fontes de Erro

- Agora vamos verificar qual é o valor esperado da função de custo em relação a um valor y
  - Vamos definir um modelo médio em relação a todos os possíveis *datasets* $f_\mu[x] = \mathbb{E}_D\left[f\left[x, \theta[D]\right]\right]$

$$\mathbb{E}_D\left[\mathbb{E}_y\left[L[x]\right]\right] = \boxed{\mathbb{E}_D\left[\left(f[x,\theta] - f_\mu[x]\right)^2\right]} + \left(f_\mu[x] - \mu[x]\right)^2 + \sigma^2$$

Quanto o modelo atual varia em relação ao modelo médio

Variância

Fontes de Erro

- Agora vamos verificar qual é o valor esperado da função de custo em relação a um valor y
  - Vamos definir um modelo médio em relação a todos os possíveis *datasets* $f_\mu[x] = \mathbb{E}_D\left[f\left[x, \theta[D]\right]\right]$

$$\mathbb{E}_D\left[\mathbb{E}_y[L[x]]\right] = \mathbb{E}_D\left[\left(f[x,\theta] - f_\mu[x]\right)^2\right] + \boxed{\left(f_\mu[x] - \mu[x]\right)^2} + \sigma^2$$

Quanto o modelo médio difere da média real da variável alvo

Viés

Fontes de Erro

- Agora vamos verificar qual é o valor esperado da função de custo em relação a um valor y
  - Vamos definir um modelo médio em relação a todos os possíveis *datasets* $f_\mu[x] = \mathbb{E}_D\left[f\left[x, \theta[D]\right]\right]$

$$\mathbb{E}_D\left[\mathbb{E}_y\left[L[x]\right]\right] = \mathbb{E}_D\left[\left(f[x,\theta] - f_\mu[x]\right)^2\right] + \left(f_\mu[x] - \mu[x]\right)^2 + \boxed{\sigma^2}$$

Quanto o modelo médio difere da média real da variável alvo

ruído

Fontes de Erro

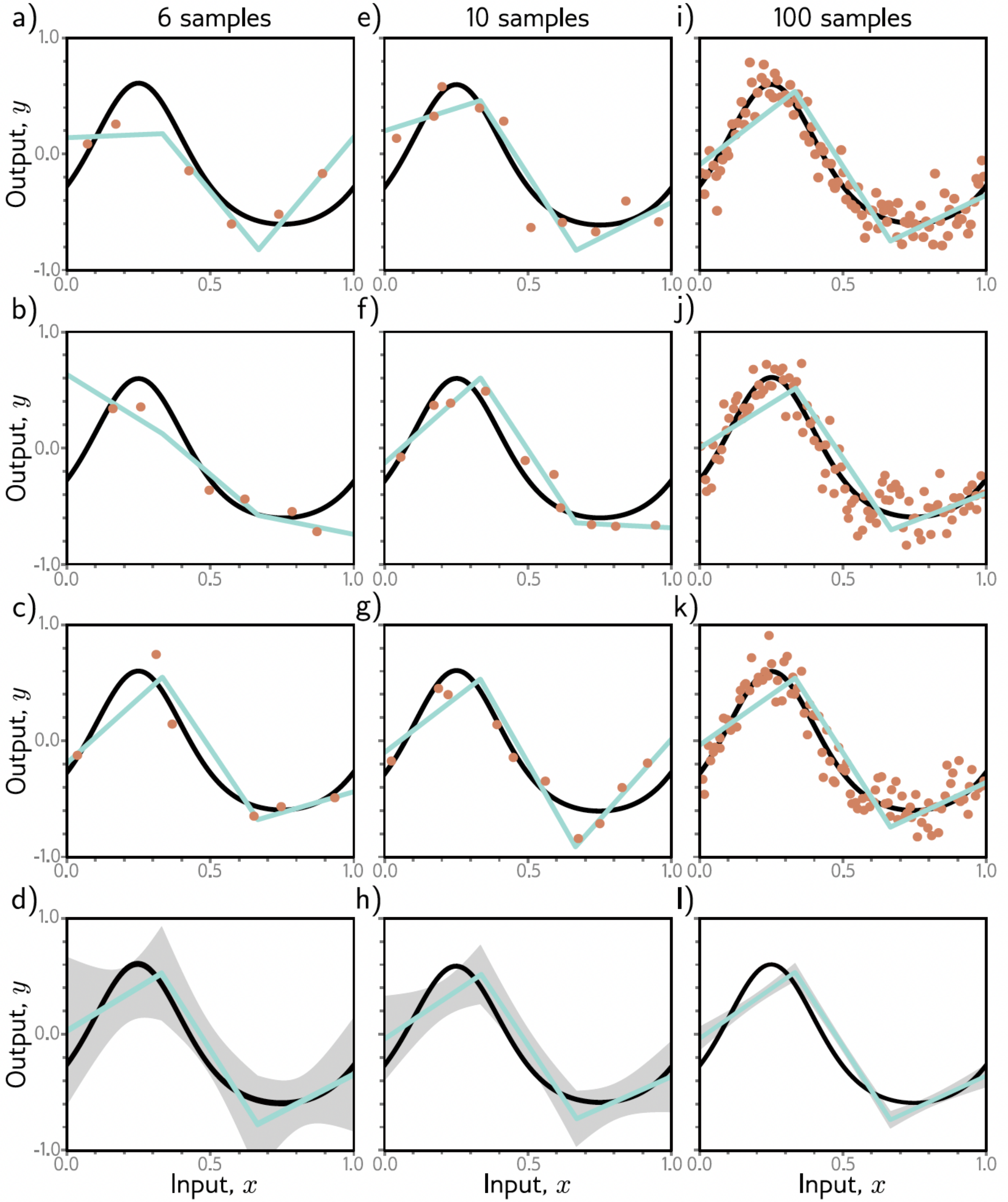
- Agora vamos verificar qual é o valor esperado da função de custo em relação a um valor y

$$\mathbb{E}_D\left[\mathbb{E}_y[L[x]]\right] = \underbrace{\boxed{\mathbb{E}_D\left[\left(f[x,\theta] - f_\mu[x]\right)^2\right]}}_{\text{Variância}} + \underbrace{\boxed{\left(f_\mu[x] - \mu[x]\right)^2}}_{\text{Viés}} + \underbrace{\boxed{\sigma^2}}_{\text{ruído}}$$

Em modelos lineares essas fontes de erro são três componentes aditivos, em modelos não lineares essa interação pode ser mais complexa

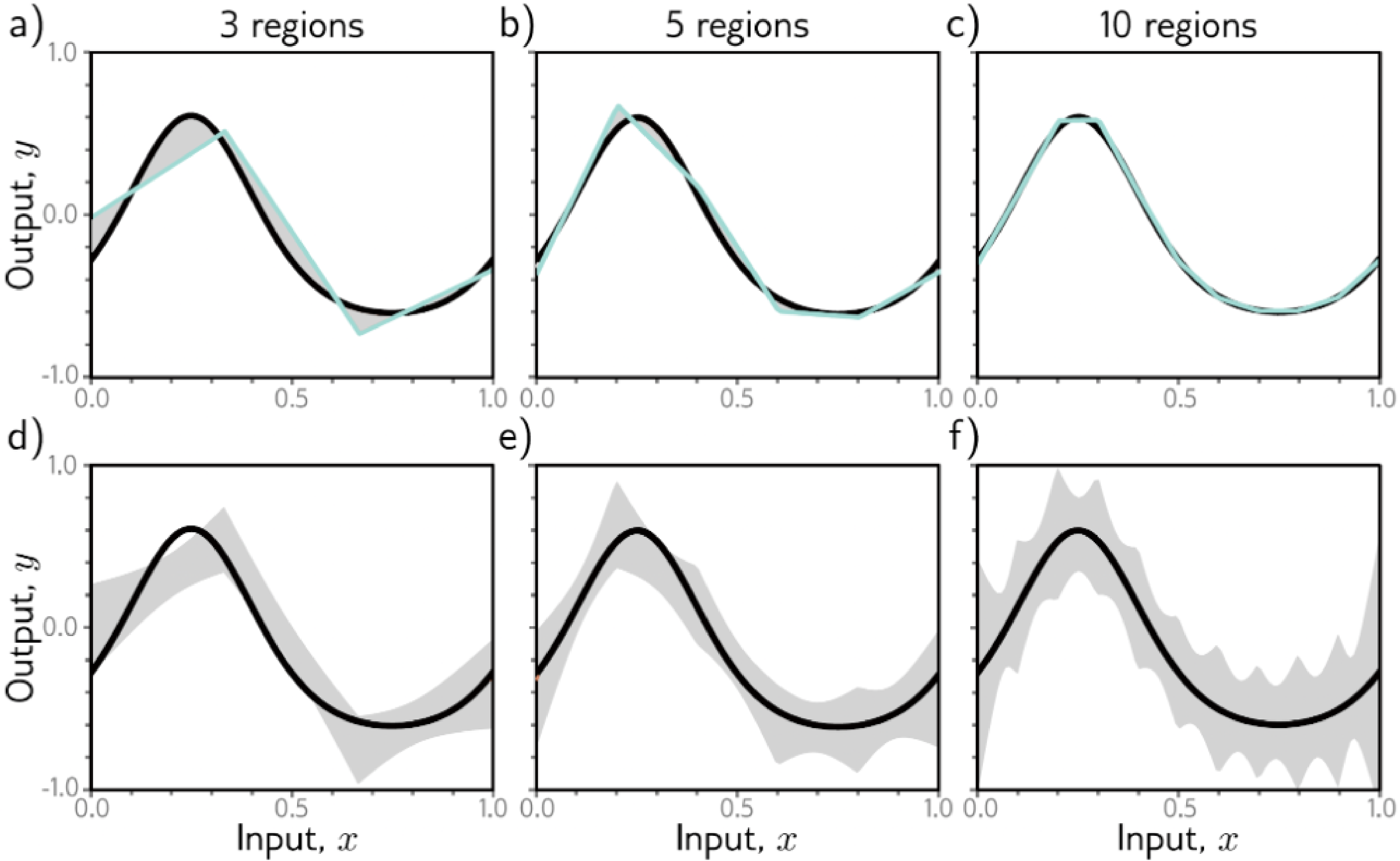
Dataset Size vs *Variance*

- Lembre-se que estamos falando do modelo que tem capacidade de modelar 3 regiões lineares
- Grande variância para 6 instâncias
- Pequena variância para 100 instâncias



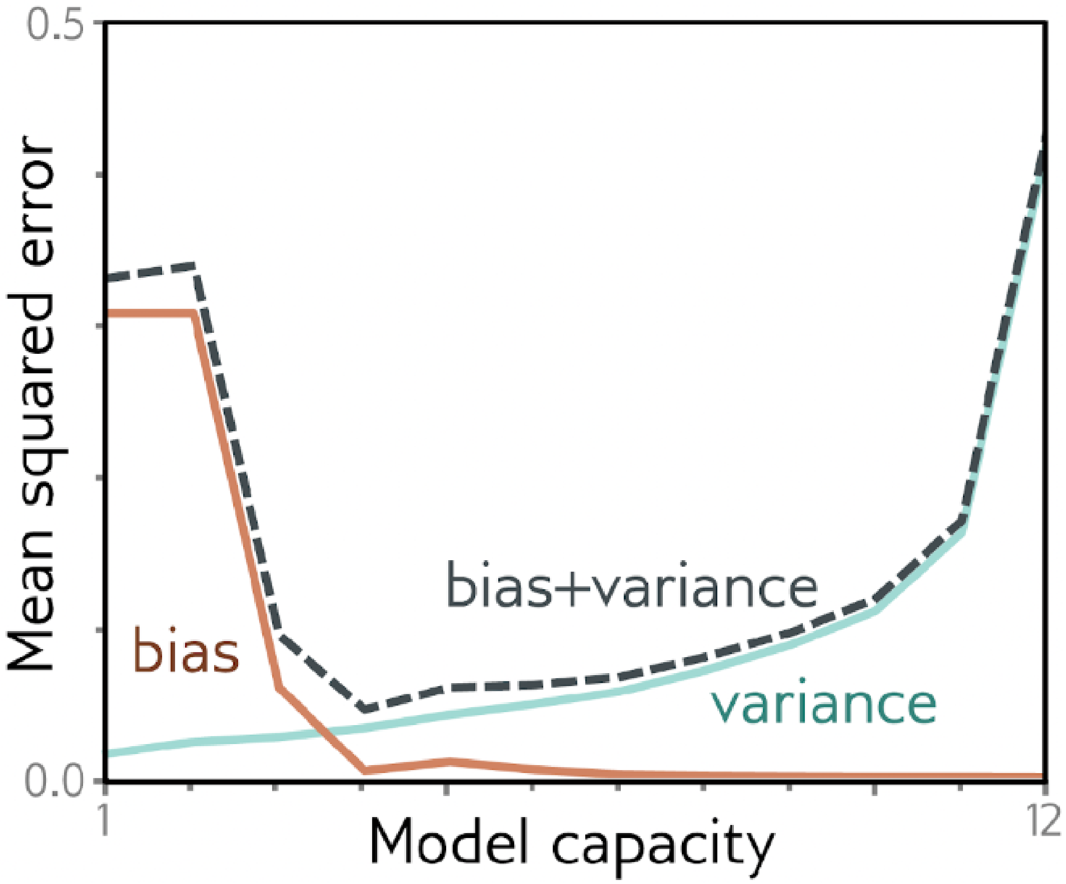
Capacidade do Modelo vs Variância

- + capacidade
- +variância



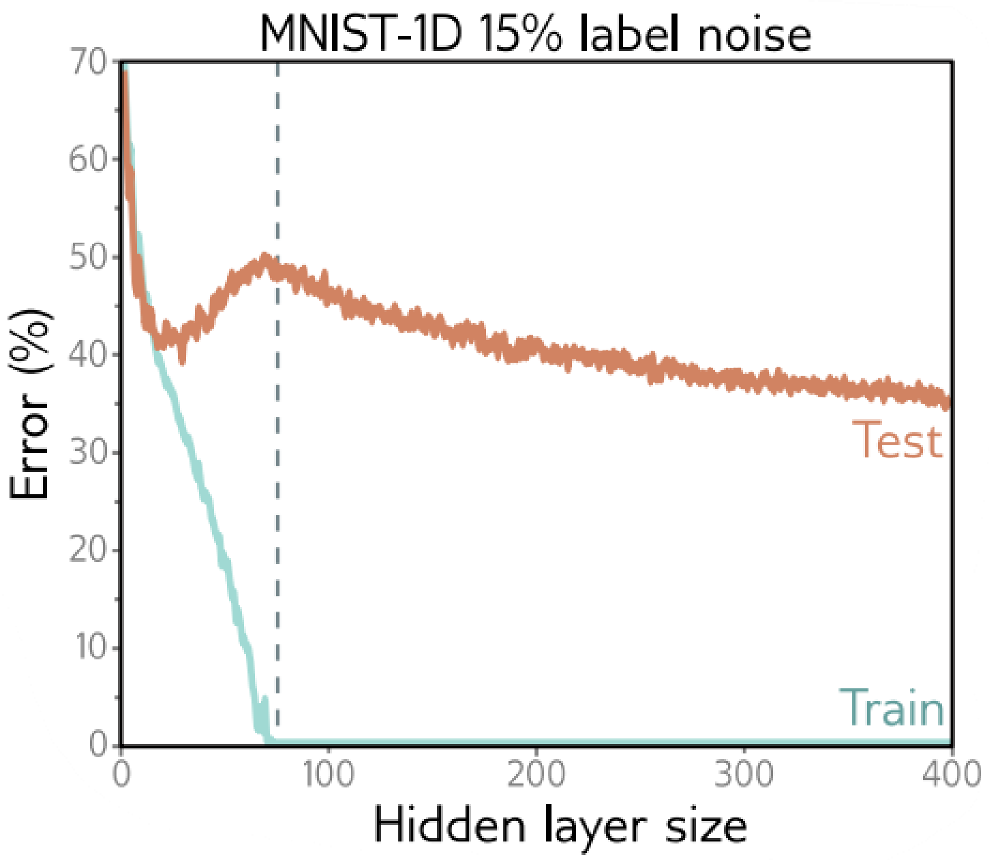
Trade-off Viés-Variância

- **Usualmente**
  - Quanto maior a capacidade, maior o erro por variância
  - Quanto menor a capacidade, maior o erro por viés



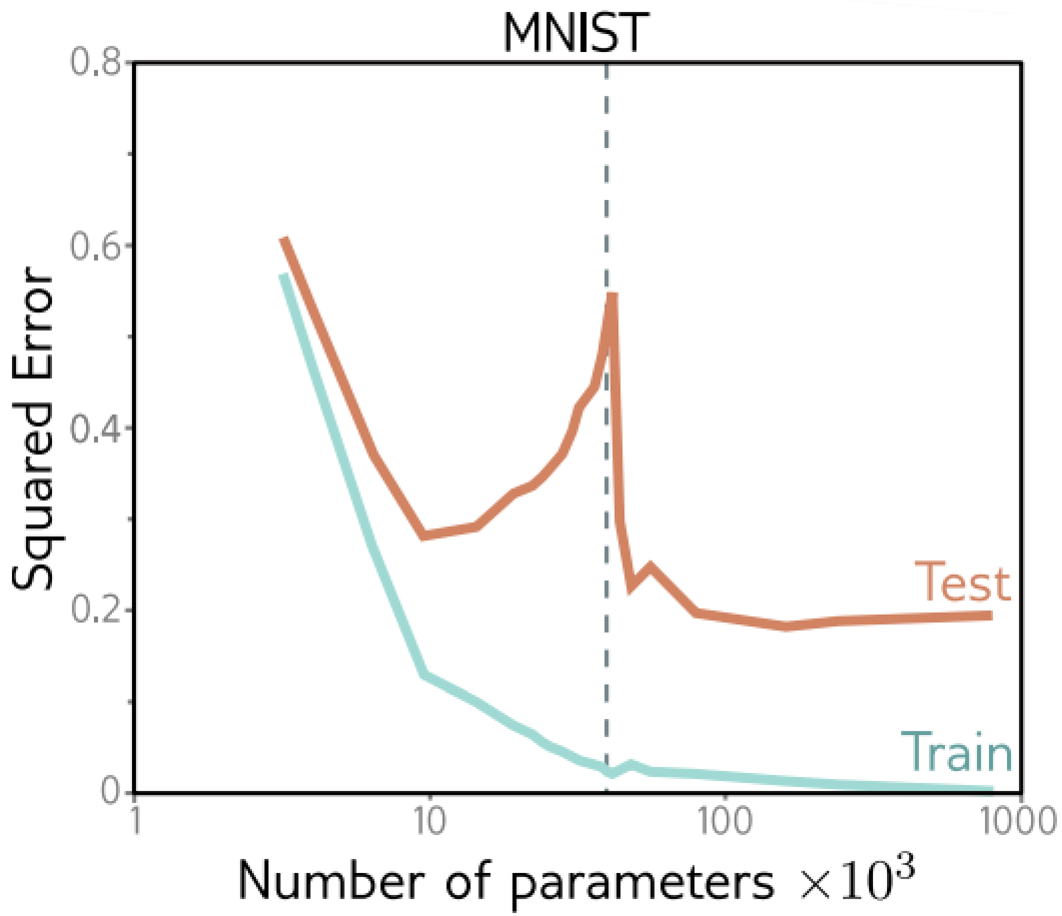
Double Descent

- Será que temos esse trade-off com redes neurais?
  - Observe o comportamento no gráfico
    - MNIST-1D



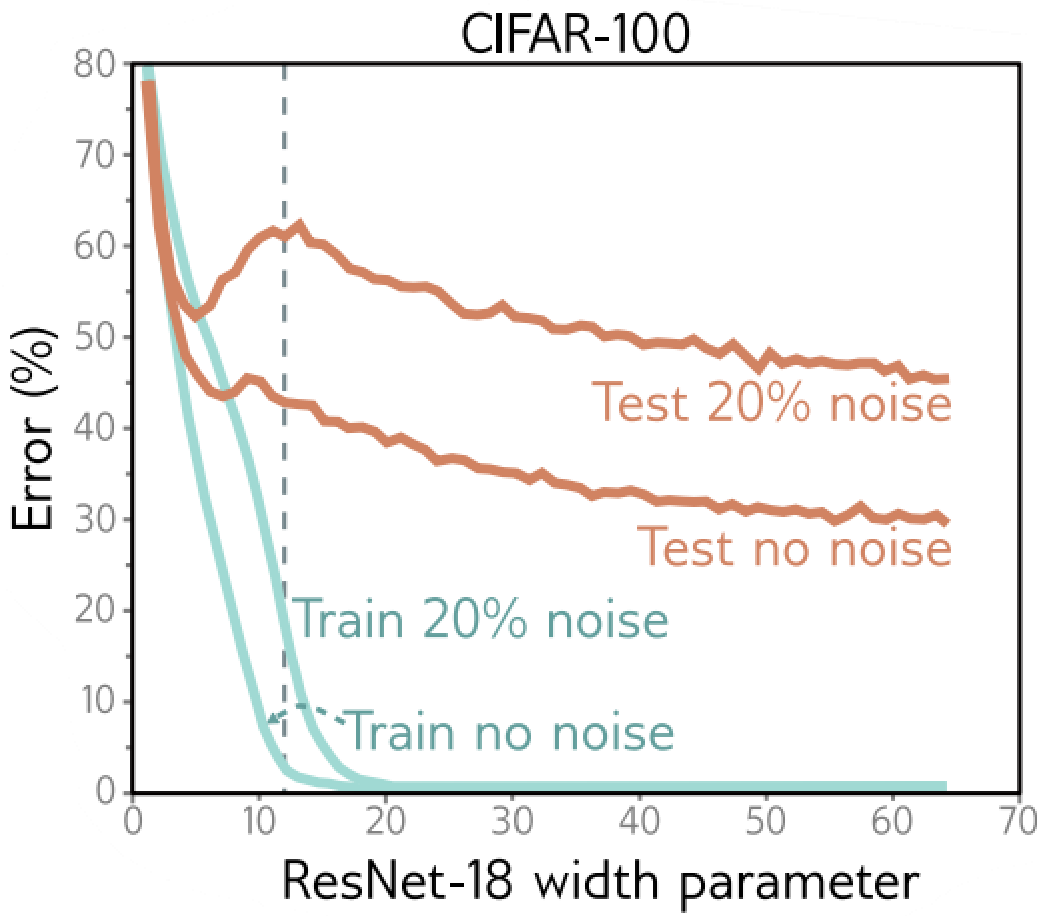
Double Descent

- Será que temos esse trade-off com redes neurais?
  - Observe o comportamento no gráfico
    - MNIST-1D
    - MNIST



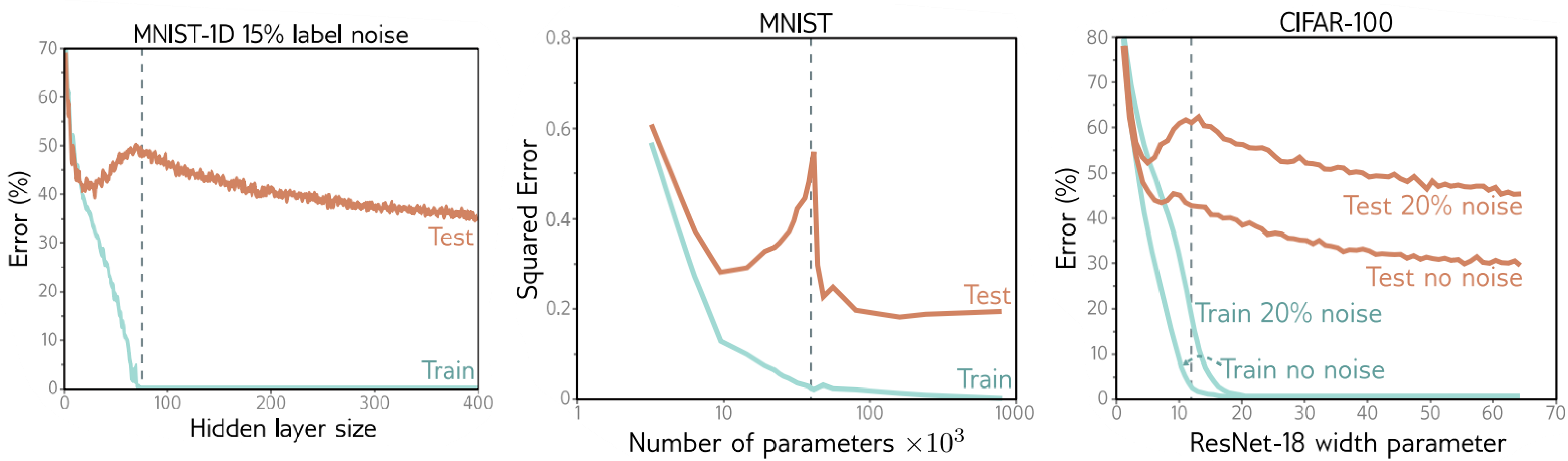
Double Descent

- Será que temos esse trade-off com redes neurais?
  - Observe o comportamento no gráfico
    - MNIST-1D
    - MNIST
    - CIFAR-100



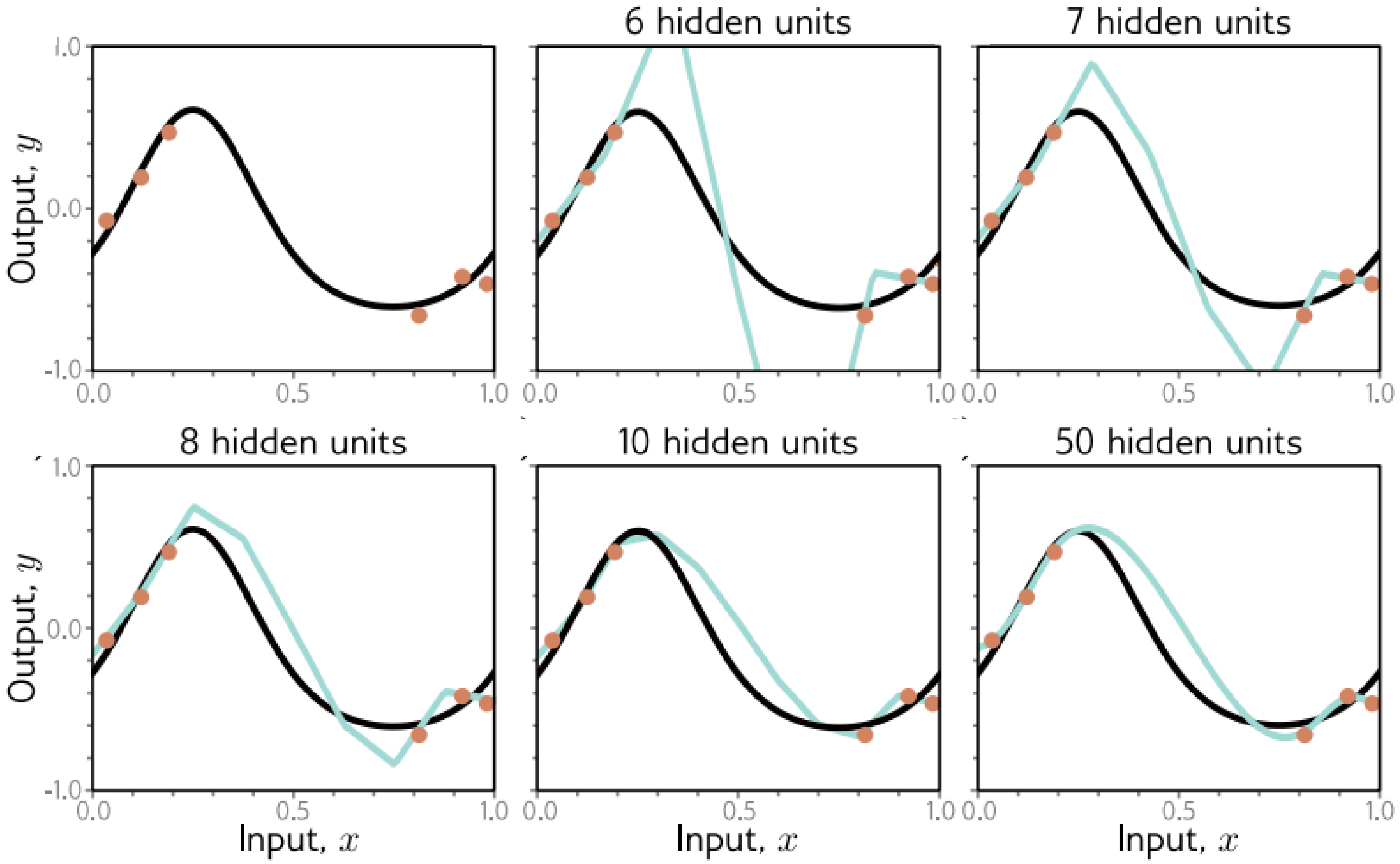
Double Descent

- A linha pontilhada marca o local onde o número de parâmetros da rede é igual ao número de instâncias de treino



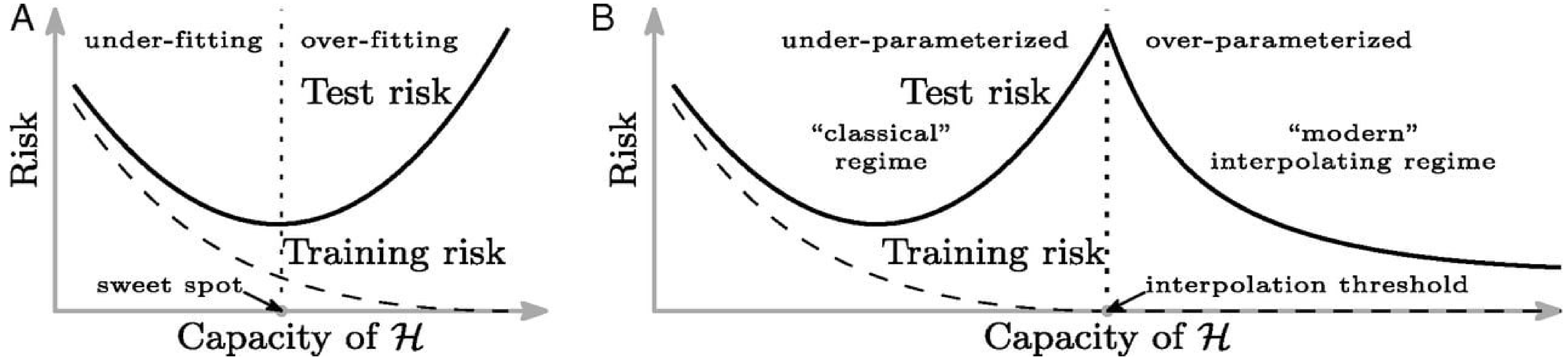
Double Descent

- Um outro experimento para ilustrar



Double Descent

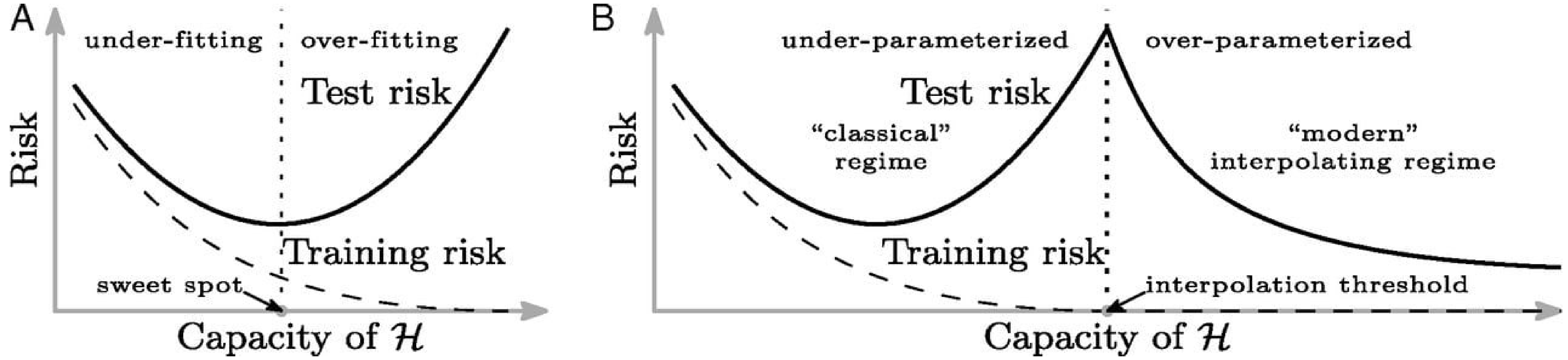
- Por que esse fenômeno ocorre?



BELKIN, Mikhail et al. Reconciling modern machine-learning practice and the classical bias–variance trade-off. **Proceedings of the National Academy of Sciences**, v. 116, n. 32, p. 15849-15854, 2019.

Double Descent

- Talvez a inicialização encoraje funções mais suaves
- Talvez o treinamento encoraje funções mais suaves
- Ninguém sabe ao certo, são apenas bons palpites ☺



BELKIN, Mikhail et al. Reconciling modern machine-learning practice and the classical bias–variance trade-off. **Proceedings of the National Academy of Sciences**, v. 116, n. 32, p. 15849-15854, 2019.

Regularização

- L1 e L2
  - Enviesamos nossa função de custo para dar preferência para parâmetros de baixa magnitude
  - Do ponto de vista probabilístico, estamos adicionando um *prior* no critério de máxima verossimilhança
    - Como redes neurais são atuam em problemas gerais, nosso *prior* tem que ser algo “genérico”

$$\hat{\theta} = arg\min_{\theta}\left[\sum_{i=1}^{N} l\left(f\left(x^{(i)},\theta\right),y^{(i)}\right) + \lambda\|\theta\|^2\right]$$

$$\theta_{t+1} = \theta - \eta \frac{\partial\left(\sum_{i=1}^{N} l\left(f\left(x^{(i)}, \theta\right), y^{(i)}\right) + \lambda\|\theta\|^2\right)}{\partial\theta}$$

$$\theta_{t+1} = \theta - \eta \left(\frac{\partial\left(\sum_{i=1}^{N} l\left(f\left(x^{(i)}, \theta\right), y^{(i)}\right)\right)}{\partial\theta} + \frac{\partial(\lambda\|\theta\|^2)}{\partial\theta} \right)$$

$$\theta_{t+1} = \theta - \eta \left(\frac{\partial\left(\sum_{i=1}^{N} l\left(f\left(x^{(i)}, \theta\right), y^{(i)}\right)\right)}{\partial\theta} + 2\lambda\theta \right)$$

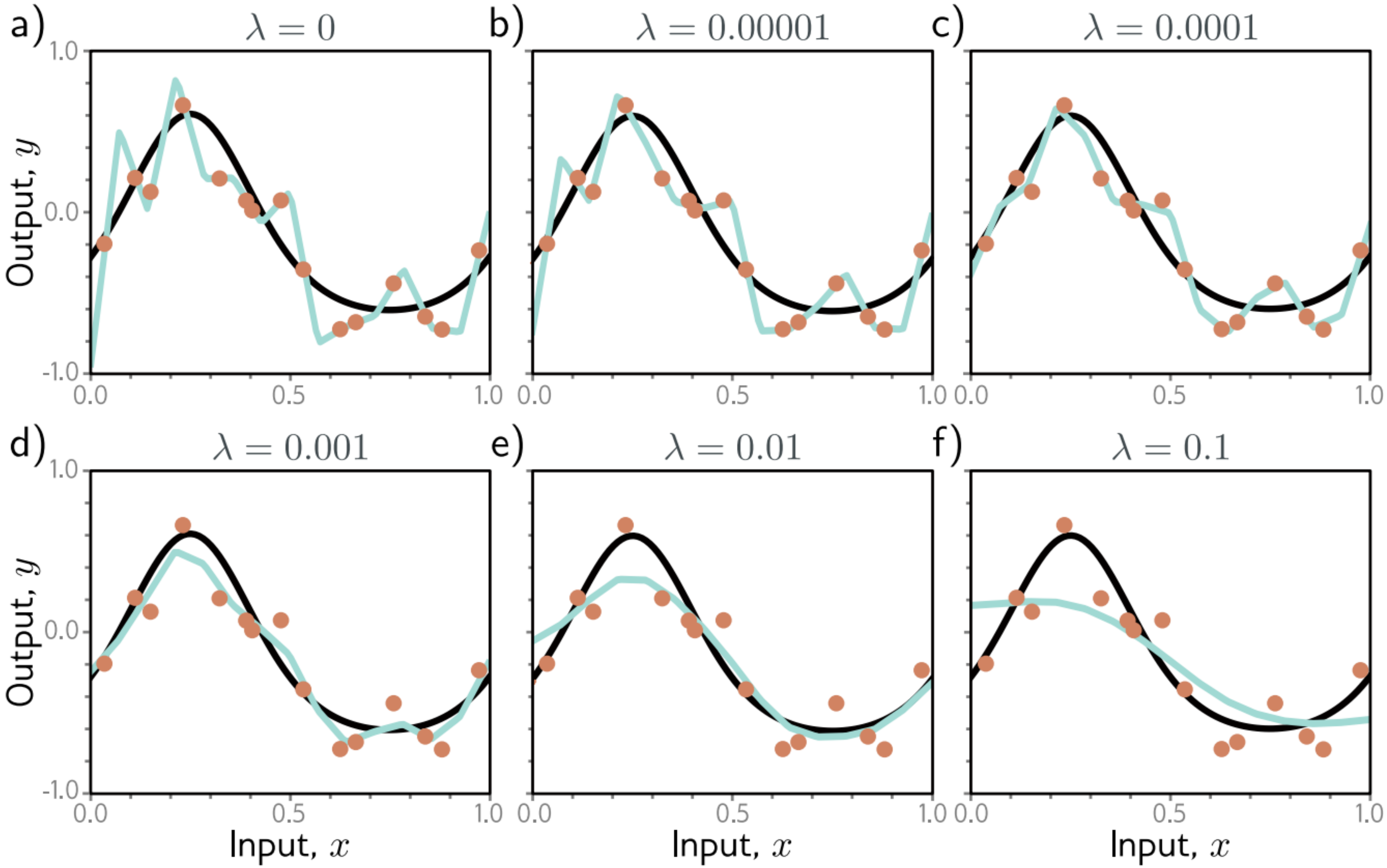
$$\theta_{t+1} = \theta - \eta \frac{\partial\left(\sum_{i=1}^{N} l\left(f\left(x^{(i)}, \theta\right), y^{(i)}\right)\right)}{\partial\theta} - \eta 2\lambda\theta$$

$$\theta_{t+1} = \theta(1 - 2\eta\lambda) - \eta \frac{\partial\left(\sum_{i=1}^{N} l\left(f\left(x^{(i)}, \theta\right), y^{(i)}\right)\right)}{\partial\theta}$$

$$\theta_{t+1} = \theta(1 - 2\eta\lambda) - \eta\nabla J$$

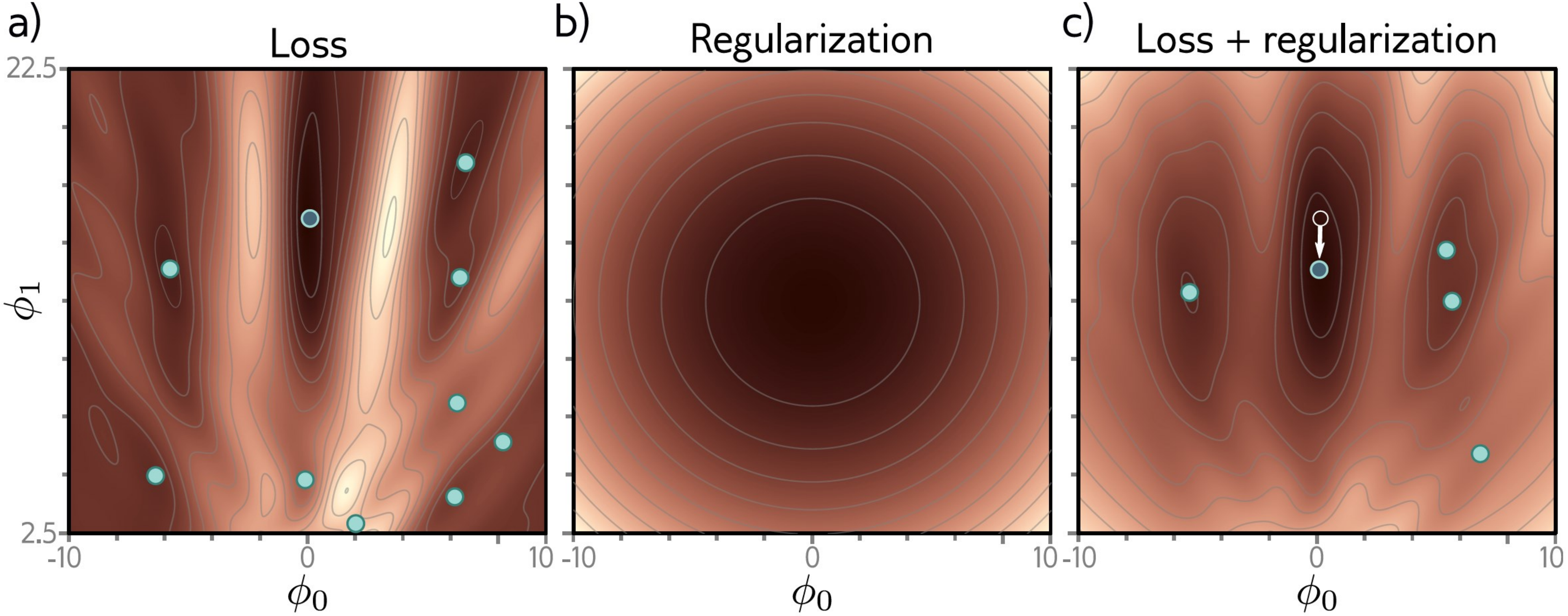
Regularização

- L1 e L2
  - Perceba ao lado o efeito de aumentar o parâmetro de regularização



Regularização

- L1 e L2

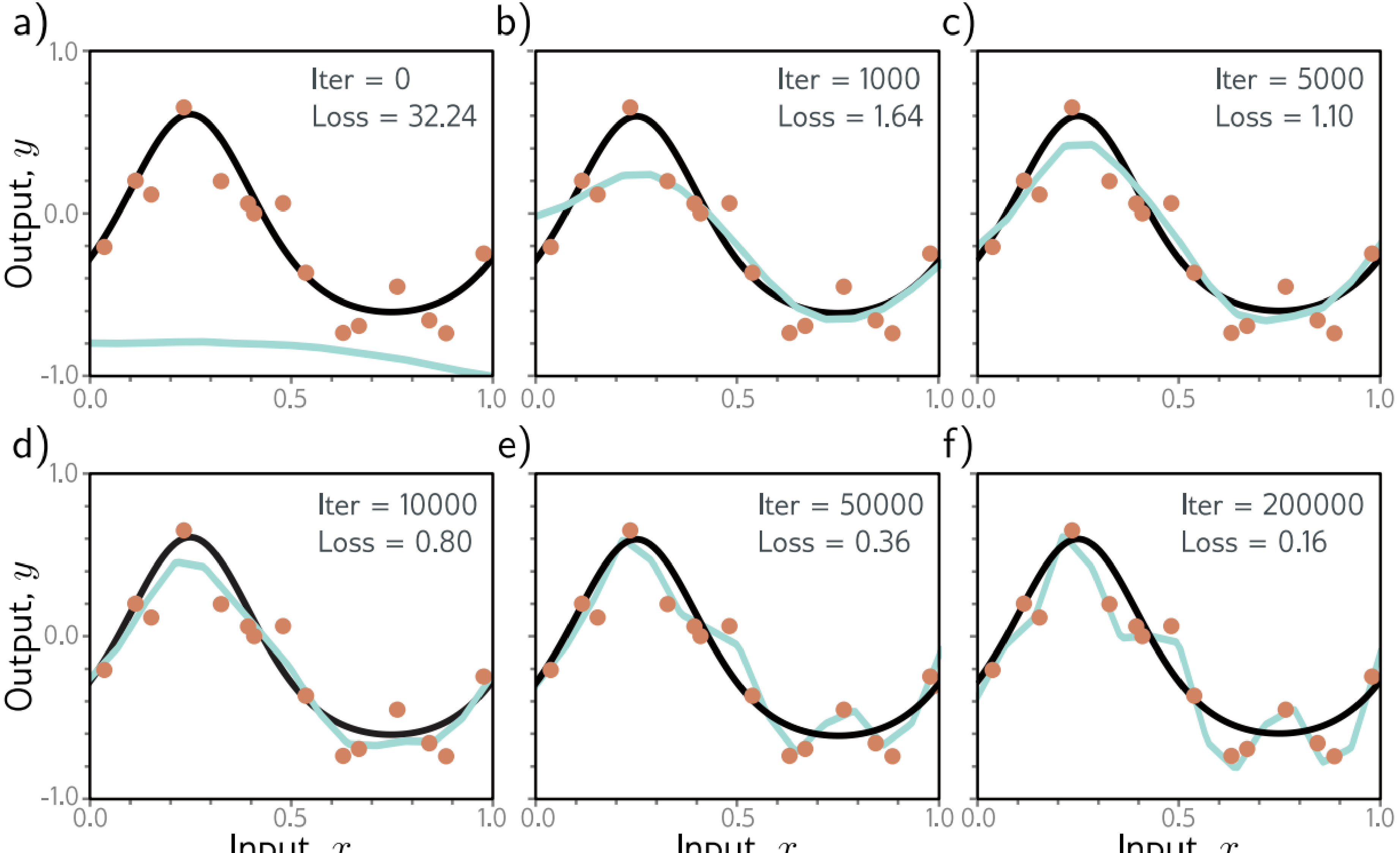


Regularização

- *Early Stopping*
  - Consiste em monitorar o desempenho da rede neural em um conjunto de dados de validação e parar o treinamento antes da loss "convergir"
  - Tem um efeito similar a regularização L2, pois não deixa os parâmetros crescerem indiscriminadamente
  - Usualmente possui apenas um hiperparâmetro associado, que define quantas iterações vamos aguardar sem que haja melhoria na função de custo em validação

Regularização

- *Early Stopping*

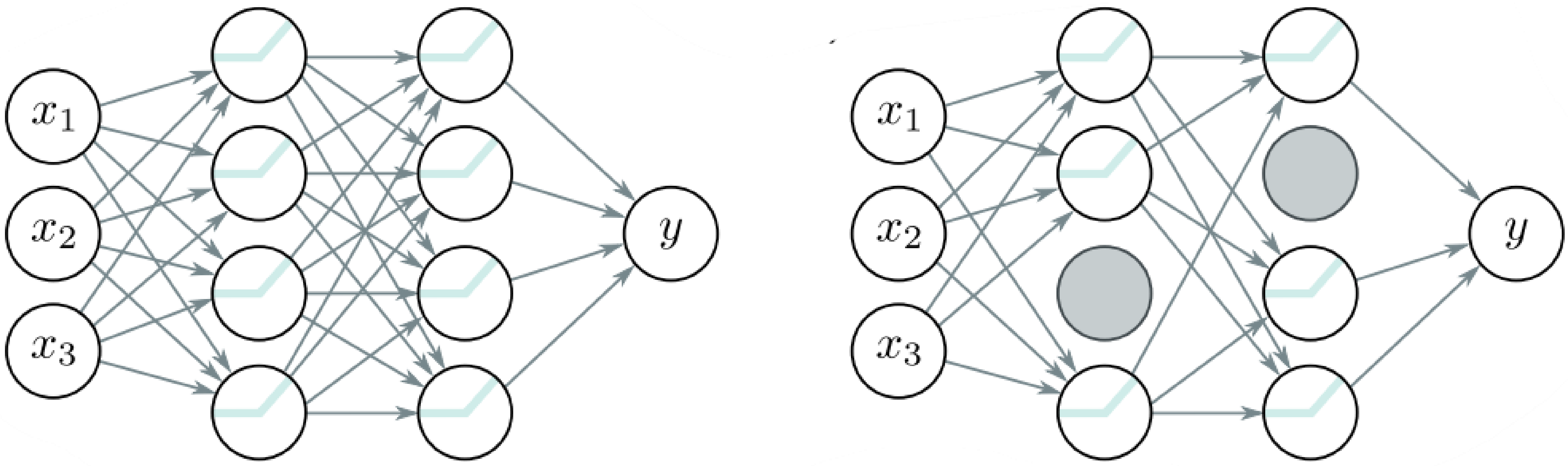


Regularização

- *Drop Out*
  - Durante o treinamento, são desligadas (ou desconectadas) de forma aleatória algumas unidades em uma camada oculta
  - Tem o efeito prático de multiplicarmos as ativações por um ruído que segue uma distribuição de Bernoulli

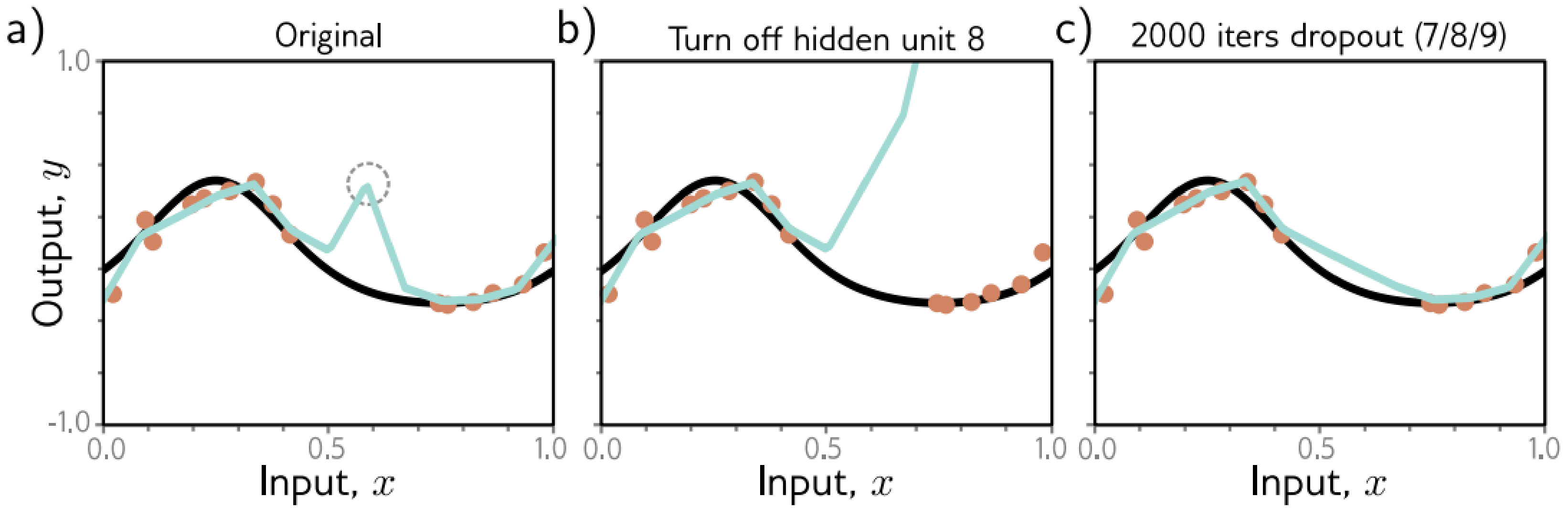
Regularização

- *Drop Out*



Regularização

- *Drop Out*
  - Perceba que na figura a) abaixo, o modelo já se ajusta perfeitamente aos dados de treinamento (continuar treinando não deve ajudar)
  - Contudo, se uma das unidades for desligada, o pico no meio da função causará um grande aumento na *loss* (ver b)
  - Se continuarmos desligando aleatoriamente as unidades, esse pico no meio da função vai acabar desaparecendo (ver c)



Regularização

- *Drop Out*
  - Na inferência temos duas opções para compensar o aumento da intensidade das ativações:
    - Multiplicamos as ativações por $1 - dropout\_rate$
    - Usar *Monte Carlo Drop Out,* rodar o *forward* da rede múltiplas vezes com diferentes unidades desligadas e combinar as diferentes saídas (de forma análoga a um *ensemble*)

Regularização

- *Data Augmentation*
  - Consiste em gerar novas instâncias fazendo manipulações nos dados
    - Rotações, Flip, Mudança de Cor, Equalização, Random Crop, Adição de Ruído,...
    - Adicionamos as invariâncias desejadas via augmentation

Dog

Dog

Regularização

- Implícita
  - É um fato peculiar que a descida de gradiente encontre soluções que generalizam para além do conjunto de treinamento
  - Observou-se que tanto a descida de gradiente em batch (N) e a descida de gradiente estocástica (1) não fazem uma trajetória direta ao mínimo local da função, existe uma preferência para algumas soluções em detrimento de outras
    - Essas preferências recebem o nome de regularização implícita

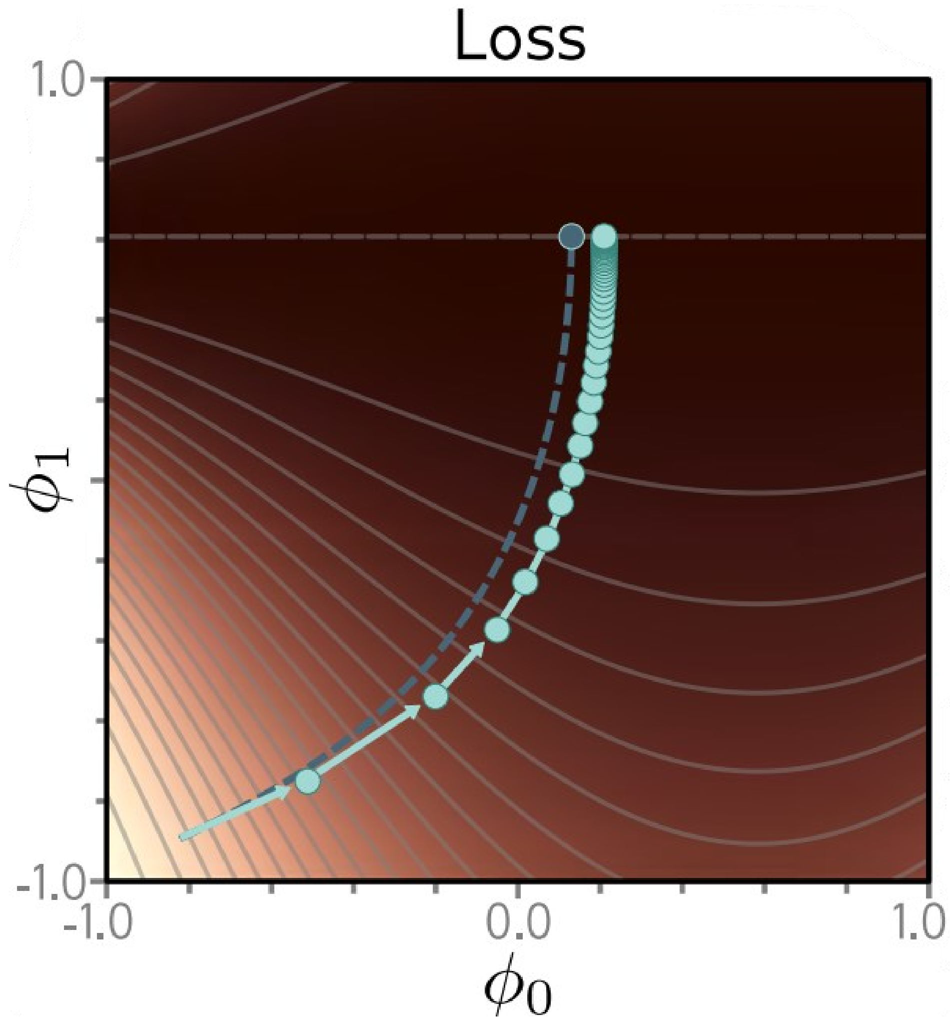
Regularização Implícita

- A descida de gradiente contínua pode ser representada a partir da seguinte equação diferencial

$$\frac{\partial \theta}{\partial t} = -\frac{\partial L}{\partial \theta}$$

- Contudo, quando implementamos a descida de gradiente usamos a versão discreta

$$\theta_{t+1} = \theta_t - \eta \frac{\partial L[\theta_t]}{\partial \theta}$$



Regularização Implícita

- Foi demonstrado que a versão discreta da descida de gradiente aproxima a versão contínua com a seguinte função de custo

$$L_{GD}(\theta) = L(\theta) + \frac{\eta}{4}\left\|\frac{\partial L}{\partial \theta}\right\|^2$$

De modo que a solução é repelida de lugar nos quais a segunda norma do gradiente é grande (alta declividade)

BARRETT, David GT; DHERIN, Benoit. Implicit gradient regularization. **arXiv preprint arXiv:2009.11162**, 2020.

Regularização Implícita

- Foi demonstrado que a versão discreta da descida de gradiente estocástica aproxima a versão contínua com a seguinte função de custo

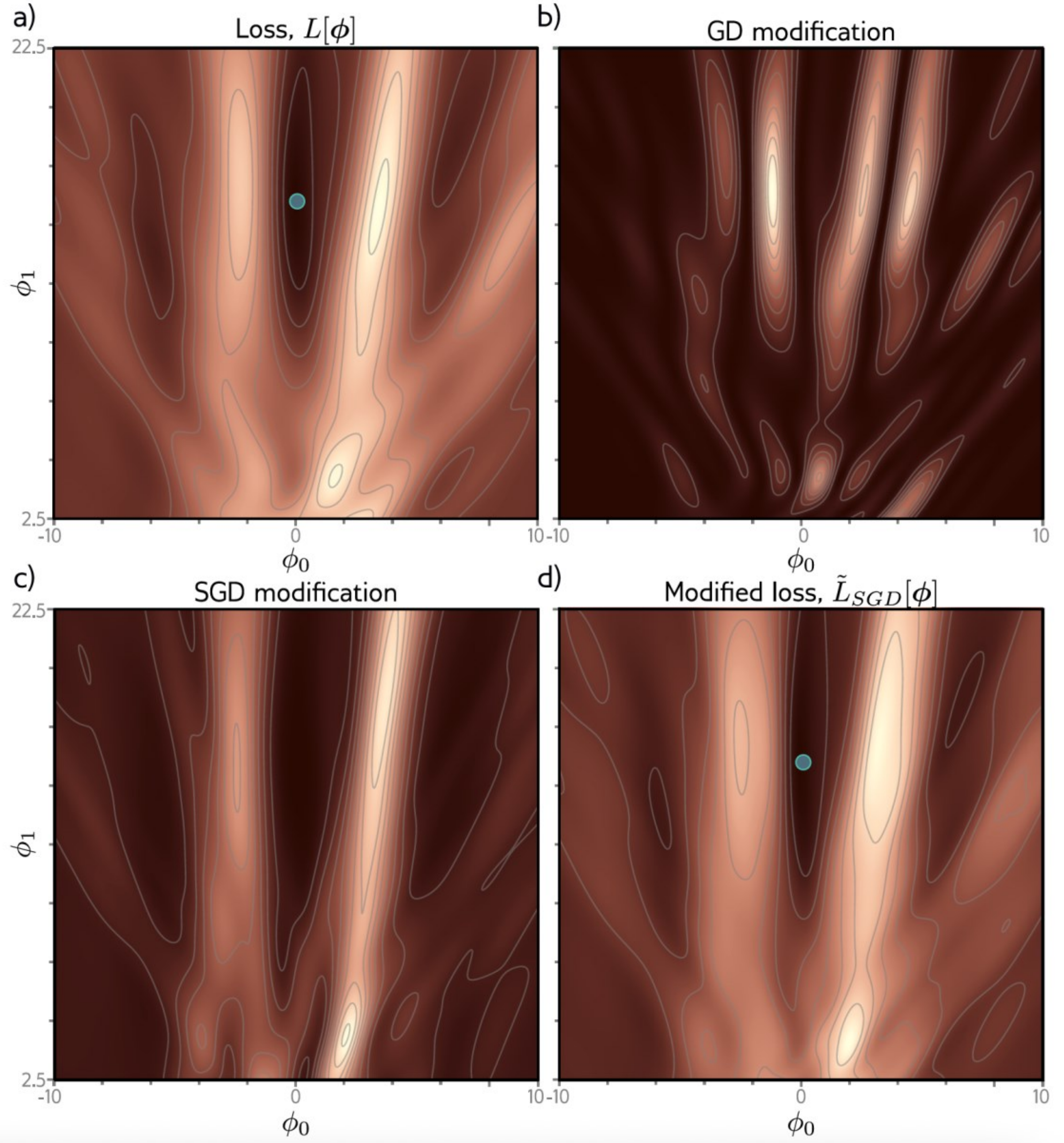
$$L_{SGD}(\theta) = L(\theta) + \frac{\eta}{4}\left\|\frac{\partial L}{\partial \theta}\right\|^2 + \frac{\eta}{4B}\sum_{b=1}^{B}\left\|\frac{\partial L_b}{\partial \theta} - \frac{\partial L}{\partial \theta}\right\|^2$$

- De modo que a solução é repelida de lugar nos quais a segunda norma do gradiente é grande (alta declividade)
- O segundo termo de regularização corresponde a variância do gradiente nos diferentes batchs
  - Em outras palavras, SGD favorece regiões cujo gradiente é estável

SMITH, Samuel L. et al. On the origin of implicit regularization in stochastic gradient descent. **arXiv preprint arXiv:2101.12176**, 2021.

Regularização Implícita

- Ambas modificações não mudam as posições dos mínimos
  - Mas podem mudar as trajetórias e qual mínimo local encontrado



Referências:

- Sugere-se ***fortemente*** a leitura de:
  - Capítulo 9 de Understanding Deep Learning
    - https://udlbook.github.io/udlbook/