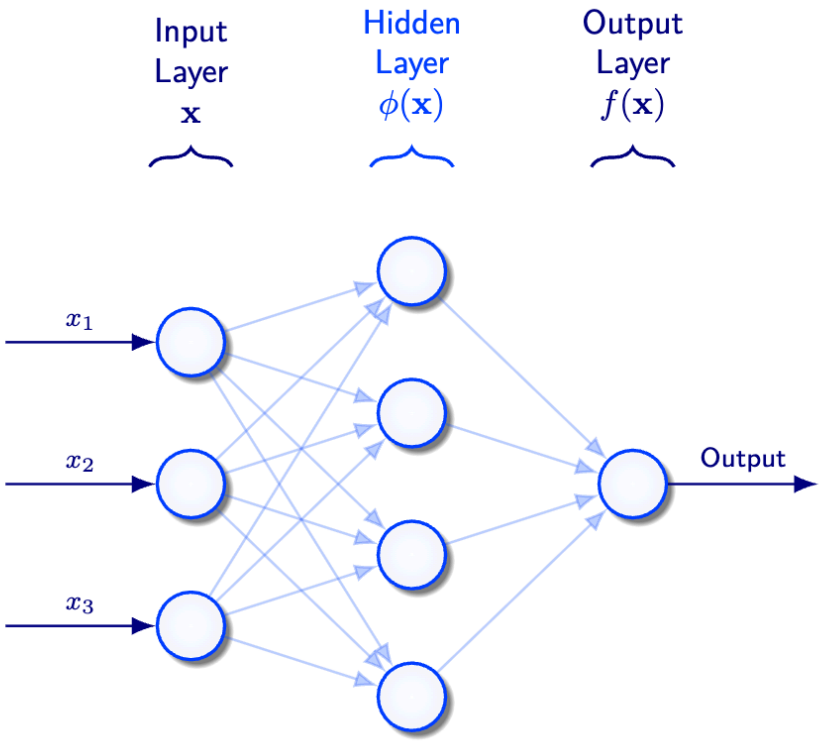
BAYESIAN NEURAL NETWORKS: AN INTRODUCTION AND SURVEY

Eduardo Azeredo

Estatítica Computacional — Agosto de 2025

REDES NEURAIS - FUNÇÃO DE ATIVAÇÃO

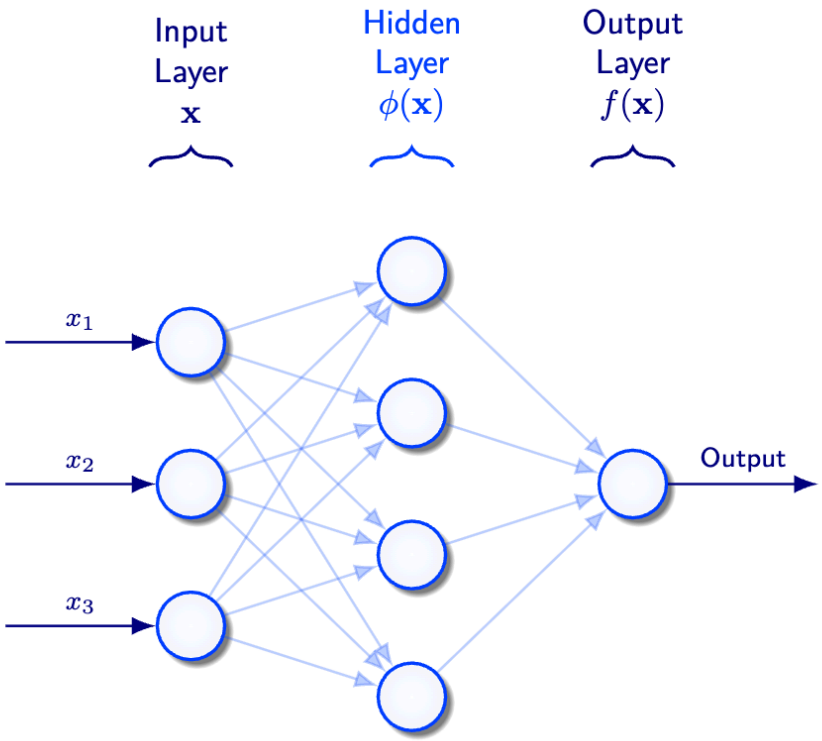


Podemos representar uma rede neural como uma função:

$$\phi_j = \sum_{i=1}^{N_1} a(x_i w_{ij}^1)$$

onde a é uma função de ativação.

REDES NEURAIS - OUTPUT



A saída da rede neural é dada por:

$$f_k = \sum_{j=1}^{N_2} g(\phi_j w_{ik}^2)$$

onde g ou é a função identidade (regressão) ou uma função softmax (classificação).

REDES NEURAIS - TREINAMENTO

- O treinamento da rede consiste em achar os pesos $\mathbf{w}$ que melhor aproximam a função desejada;
- A minimização é feita via descida de gradiente, calculando os gradiente em relação aos pesos;
- Esse processo é conhecido como backpropagation.
- Em termos estatísticos, os pesos $\mathbf{w}$ são os parâmetros desconhecidos do modelo, e o treinamento é o processo de estimá-los. **Esta é uma abordagem frequentista**.

HISTÓRIA

- Tishby, Levin e Solla (1989), ao estudar as propriedades estatísticas de NN, mostraram que atribuir um prior aos pesos pode ser usado para gerar uma posterior, sem indicar como obtê-la;
- Denker e LeCun (1991) propõem um método usando aproximação de Laplace para calcular a posterior;
- Hinton e Van Camp (1993) propõem o uso de Inferência Variacional (VI) para aproximar a posterior;
- Neal (1996) propõe o uso de MCMC para amostrar a posterior usando Hybrid/Hamiltonian Monte Carlo (HMC);

REDES NEURAIS BAYESIANAS - INTRODUÇÃO

- Se atribuirmos que os pesos $\mathbf{w}$ são variáveis aleatórias, e que podemos atribuir uma distribuição a eles, temos uma rede neural bayesiana;
- Vale então o Teorema de Bayes de que

$$\pi(\mathbf{w}|\mathcal{D}) = \frac{p(\mathbf{w})p(\mathcal{D}|\mathbf{w})}{\int p(\mathbf{w})p(\mathcal{D}|\mathbf{w})d\mathbf{w}} = \frac{p(\mathbf{w})p(\mathcal{D}|\mathbf{w})}{p(\mathcal{D})}$$

- Onde $\mathcal{D}$ é o conjunto de dados, $p(\mathbf{w})$ é a distribuição a priori dos pesos, $p(\mathcal{D}|\mathbf{w})$ é a verossimilhança dos dados, $\pi(\mathbf{w}|\mathcal{D})$ é a distribuição posterior verdadeira dos pesos, e $p(\mathcal{D})$ é a evidência.

- A predição desse modelo é dada por:

$$\mathbb{E}_{\pi}(\mathbf{w}) = \int f(\mathbf{w})\pi(\mathbf{w}|\mathcal{D})d\mathbf{w}$$

- "opposed to an optimisation scheme used in the frequentist setting [...] our predictions are represented in the form of valid conditional probabilities";
- O problema é que, mesmo para problemas simples, não conseguimos calcular a integral. Para isso usamos métodos de aproximação, sendo o mais comum VI, ou Variational Inference.

INTRODUÇÃO

- A ideia central da VI é transformar o problema de inferência em um problema de otimização;
- Assumimos uma distribuição para a densidade $q_\theta(\mathbf{w})$ e através usando divergência Kullback-Leibler (KL) aproximamos a posterior verdadeira $\pi(\mathbf{w}|\mathcal{D})$; Definimos primeiro divergência KL

$$KL[q_\theta(\mathbf{w})||p(\mathbf{w}|\mathcal{D})] = \int q_\theta(\mathbf{w}) \log \left(\frac{q_\theta(\mathbf{w})}{p(\mathbf{w}|\mathcal{D})} \right) d\mathbf{w}$$

- Queremos então minimizar a divergência KL em relação a θ.

ELBO

Podemos expandir a expressão da divergência KL como

$$\begin{aligned} KL[q_\theta(\mathbf{w})||p(\mathbf{w}|\mathcal{D})] &= \mathbb{E}_q\left(\log\frac{q_\theta(\mathbf{w})}{p(\mathbf{w})} - \log p(\mathbf{w}|\mathcal{D})\right) + \log p(\mathcal{D}) \\ &= KL\left(q_\theta(\mathbf{w})||p(\mathbf{w})\right) - \mathbb{E}_q\left[\log p(\mathcal{D}|\mathbf{w})\right] + \log p(\mathcal{D}) \\ &= -\mathcal{F}[q_\theta] + \log p(\mathcal{D}) \end{aligned}$$

Onde $\mathcal{F}[q_\theta] = -KL\left(q_\theta(\mathbf{w})||p(\mathbf{w})\right) + \mathbb{E}_q\left[\log p(\mathcal{D}|\mathbf{w})\right]$, que depende somente de componentes conhecidos. Não somente, o termo $\log p(\mathcal{D})$ é constante em relação a θ, e pode ser ignorado na otimização. Temos então o novo problema de otimização:

$$\theta^* = \arg\max_\theta -\mathcal{F}[q_\theta]$$

LIMITAÇÕES

Até então os resultados apresentados se restringem a uma NN com apenas uma camada escondida e possuem certas desvantagens:

- A distribuição tem que ser possível de ser fatorizada entre os pesos o que sacrifica a capacidade dos pesos de demonstrar correlações;
- Barber e Bishop (1998) estendem o domínio do problema para permitir correlações usando uma distribuição gaussiana multivariada, mas isso aumenta o número de parâmetros a serem otimizados;
- A função de ativação deve ser sigmóide que tem um gradiente que pode resultar em baixa performance no aprendizado.

SOLUÇÕES

- Welling e Teh (2011) propões o uso de mini-batches e SGD para otimizar a ELBO;
- Graves (2011) propõem o uso de MFVB de uma forma que a função $\mathcal{F}[q_\theta]$ como a soma de duas esperanças que podem ser resolvidas via integração Monte Carlo para achar os gradientes;

$$\mathcal{F}[q_\theta] = \mathbb{E}_q[\log(p(\mathcal{D}|\mathbf{w}))] - \mathbb{E}_q[\log(p_\theta(\mathbf{w})) - \log(p(\mathbf{w}))]$$

- Kingma e Welling (2014) propõem o uso do "truque de reparametrização"para reduzir a variância da estimação dos gradientes.

DROPOUT COMO VI

Gal (2016) propõe o uso de Dropout (um método de regularização) para introduizr estocasticidade nos parâmetros de uma rede neural.

$$\mathbf{W}_\rho^1 = diag(\rho)\mathbf{W}^1$$
$$\Phi_\rho = a(\mathbf{X}^T\mathbf{W}_\rho^1)$$

onde ρ é um vetor de uma Bernoulli(p) e $diag(\rho)$ é uma matriz diagonal com os elementos de ρ na diagonal. Dessa forma preservamos alguma correlação entre os parâmetros. A posterior aproximada é então dada por uma Bernoulli multiplicada pelos pesos da rede.

INTRODUÇÃO

- Neal (1996) propõe o uso de MCMC para amostrar a posterior usando Hybrid/Hamiltonian Monte Carlo (HMC) e achar diretamente os momentos da predição;
- Originalmente da física estatística, HMC é um método que usa conceitos de mecânica hamiltoniana para propor novos estados de uma cadeia de Markov;
- HMC é um método de MCMC que usa o gradiente da função alvo para guiar a amostragem;

- HMC introduz uma variável auxiliar $\mathbf{p}$, chamada de momento/energia cinética, e define a função energia hamiltoniana como

$$H(\mathbf{w}, \mathbf{p}) = U(\mathbf{w}) + K(\mathbf{v})$$

onde $U(\mathbf{w}) = -\log\left[p(\mathbf{w})\mathcal{L}(\mathbf{w})\right]$ é a energia potencial e $K(\mathbf{v}) = \sum_{i=1}^{d} \frac{v_i^2}{2m_i}$ é a energia cinética. A distribuição canônica é então dada por

$$p(\mathbf{w}, \mathbf{v}) = \frac{1}{Z} \exp(-U(\mathbf{w})) \exp(-K(\mathbf{v}))$$

- $K(\mathbf{v})$ é normalmente assumida como tendo uma distribuição normal centrada em m.

ALGORITMO

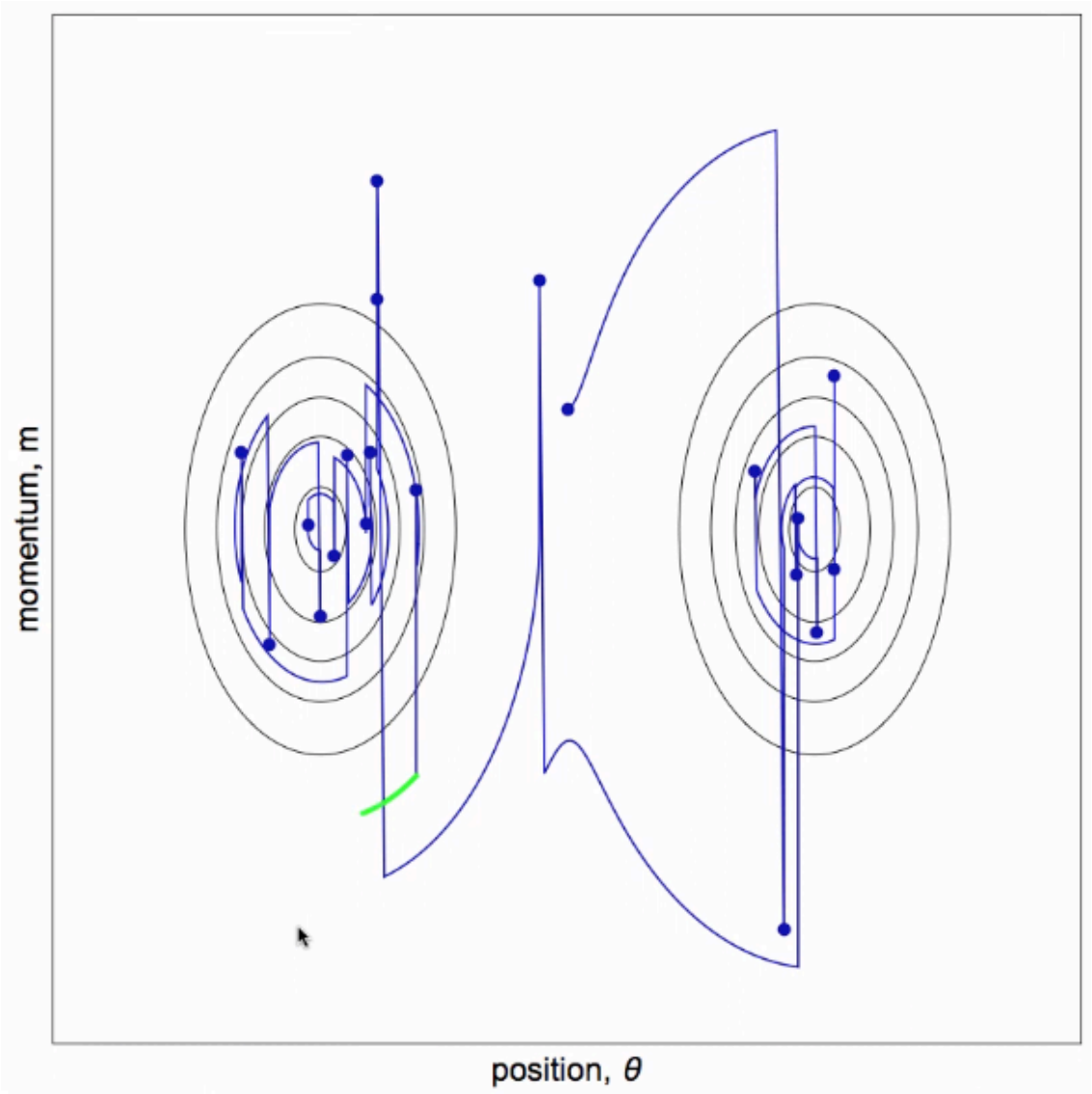
O algoritmo HMC é dado por:

- Inicializamos os pesos $\mathbf{w}$ e o momento $\mathbf{v}$;
- Amostramos da distribuição canônica de $H(\mathbf{w}, \mathbf{v})$
- Metropolis-Hastings:
  - Simulamos a dinâmica hamiltoniana por um tempo L com passos de tamanho ϵ usando o método leapfrog resultando em $(\mathbf{w}^*, \mathbf{v}^*)$;
  - Calculamos a razão r de aceitação

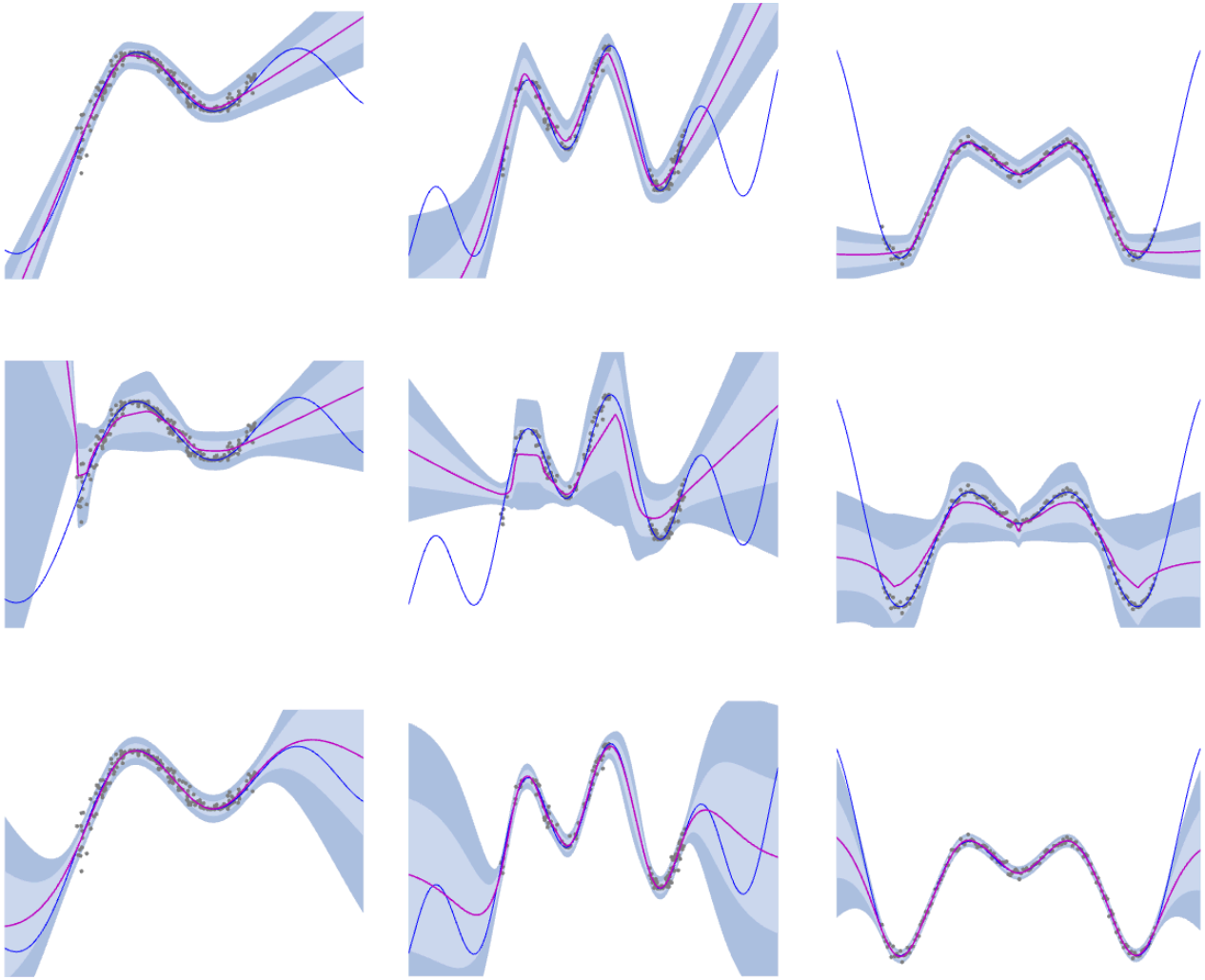
$$r = \frac{\mathcal{L}(\mathbf{w}^*)\mathbf{p}(\mathbf{w}^*)\mathcal{N}(\mathbf{v}^*)}{\mathcal{L}(\mathbf{w})\mathbf{p}(\mathbf{w})\mathcal{N}(\mathbf{v})}$$

  - Aceitamos a proposta se $u \sim U(0,1) < r$, caso contrário mantemos o estado atual.

ILUSTRAÇÃO



MÉTODOS VI



Primeira linha:
VI (BBB)

Segunda linha:
VI MC Dropout

Terceira linha:
Processo Gaussiano

CONSIDERAÇÕES FINAIS

- BNNs tendem a subestimar a variância das predições, mesmo que as médias sejam acuradas;
- Muito da pesquisa em BNNs hoje ainda depende de métodos de VI por causa da facilidade de adaptá-lo à backpropagation;
- O uso da divergência KL permite separar a verossimilhança marginal da nossa função objetivo que permite um "alvo"para a otimização, só que KL não é um medida de distância, e sim de divergência, o uso de uma distância (Hellinger) pode se mostrar vantajoso, mas a dificuldade de definir uma função objetivo viável é um desafio;
- MCMC é computacionalmente custoso e métodos de sub-sampling se mostraram ineficientes em estimar uma posterior.

REFERÊNCIAS

- GOAN, Ethan; FOOKES, Clinton. Bayesian Neural Networks: An Introduction and Survey. In: [S.l.: S.n.]. v. 2259 p. 45-87. Disponível em: `http://arxiv.org/abs/2006.12024`. Acesso em: 25 ago. 2025.
- NEAL, Radford M. MCMC using Hamiltonian dynamics. [S.l.: S.n.]. Disponível em: `https://arxiv.org/pdf/1206.1901`. Acesso em: 2 set. 2025.
- The intuition behind the Hamiltonian Monte Carlo algorithm. Ben Lambert, 15 maio 2018. Disponível em: <`https://www.youtube.com/watch?v=a-wydhEuAm0`>. Acesso em: 2 set. 2025.

OBRIGADO!

Eduardo Azeredo
Faculdade de Ciências Econômicas — UFRGS
`ufrgs.br/fce`