

Machine Learning - Assignment 11

Wei-Ju Lin

November 23, 2025

Q1. What is the impact of depth vs. width on the approximation ability of neural networks?

Is there a clear trade-off between network depth versus width in achieving approximation rates?

Under the constraint of fixed parameter count or computational budget, does there exist a relation that characterizes the optimal depth/width configuration required to minimize the error?

References:

- Eldan & Shamir (2016) — *The Power of Depth for Feedforward Neural Networks*.
Provides a theoretical proof that depth gives expressive power beyond width.
- Yarotsky (2017) — *Error bounds for approximations with deep ReLU networks*.
Establishes optimal approximation rates for deep ReLU networks and formalizes the depth-width trade-off.

Q2. In what situations should we prefer generative models vs. discriminative models?

When should we prefer generative models (to learn $p(x|y)$) over discriminative models (to learn $p(y|x)$)?

What practical criteria (such as sample size, distributional assumptions, missing data) should guide this choice?

References:

- Ng & Jordan (2001) — *On Discriminative vs. Generative Classifiers*.
Classical analysis showing that generative models perform better in small-sample regimes, while discriminative models excel with large data.
- Lasserre, Bishop & Minka (2006) — *Principled Hybrids of Generative and Discriminative Models*.
A systematic study of when and how to combine generative and discriminative modeling principles.

Q3. What factors influence the performance of softmax regression?

References:

- van den Goorbergh et al. (2022) — *The harm of class imbalance corrections for risk prediction models*.
Shows how softmax/logistic regression behaves under class imbalance and how imbalance correction affects performance.
- Fithian & Hastie (2013) — *Local case-control sampling*.
A statistical analysis of logistic/softmax regression under extreme imbalance and optimal subsampling strategies.

Q4. How is the score matching loss derived, and what is its intuition?

Where this form comes from and why minimizing the gradient difference is equivalent to learning the data distribution?

References:

- Hyvärinen (2005) — *Estimation of Non-Normalized Statistical Models by Score Matching*.
The original paper introducing score matching, including the full derivation of its loss function.

- Vincent (2011) — *A Connection Between Score Matching and Denoising Autoencoders*. Provides the most intuitive explanation: DAE training is equivalent to denoising score matching.

Q5. What is the mathematical meaning of dW_t when Brownian motion is nowhere differentiable?

In the equation $dx_t = f(x_t, t)dt + G(x_t, t)dW_t$, W_t is a Brownian motion that is nowhere differentiable. So what is dW_t from a mathematical point of view?

References:

- Särkkä & Solin (2019) — *Applied Stochastic Differential Equations*. A clean explanation of why Brownian motion is not differentiable and how dW_t should be interpreted in Itô calculus.
- Liu (2019) — *Stochastic Calculus and SDEs*. A rigorous development of the Itô integral and the formal meaning of dW_t .

Q6. Why do the Probability Flow ODE and the forward SDE share the same marginal distribution?

Why does the PF-ODE produce the same marginal density as the forward SDE even though it contains no randomness?

Is there any intuitive explanation beyond matching the Fokker-Planck equation?

References:

- Song, Sohl-Dickstein, Kingma & Ermon (2021) — *Score-Based Generative Modeling through Stochastic Differential Equations*. The primary source formalizing probability flow ODEs and proving that they share the same marginals as the corresponding SDE.
- Song & Ermon (2019) — *Generative Modeling by Estimating Gradients of the Data Distribution*. Foundational work that establishes the score-based framework, forming the basis of the PF-ODE/SDE duality.