1 Question 1: Diffusion Models - Predicting Noise vs. Matching Score

Question: Why is ”predicting noise” equivalent to ”matching the score”? Why did DDPM converge on training a model $\epsilon_\theta(x_t, t)$ to predict noise instead of directly matching the score $\nabla_x \log p(x|x_0)$?

Answer:

- **Mathematical Equivalence:** In the diffusion forward process, any state x_t is derived from x_0 via $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$. For a Gaussian distribution, the score function is:

$$\nabla_{x_t} \log p(x_t|x_0) = -\frac{x_t - \sqrt{\bar{\alpha}_t}x_0}{1-\bar{\alpha}_t} = -\frac{\epsilon}{\sqrt{1-\bar{\alpha}_t}}$$

  Thus, the score and the noise ϵ differ only by a scaling factor. Learning to predict ϵ is intrinsically learning the score function.

- **Training Stability:** While theoretically equivalent, directly predicting the score can result in exploding gradients when t is small (low noise), as the denominator approaches zero. Parameterizing the loss to predict ϵ acts as a ”weighted score matching,” which balances the gradients across different noise levels t, leading to more stable training and better sample quality.

References:

- Ho, J., et al. (2020). *Denoising Diffusion Probabilistic Models.*
- Song, Y., & Ermon, S. (2019). *Generative Modeling by Estimating Gradients of the Data Distribution.*

2 Question 2: Ito Integral - Choice of Endpoint

Question: The Ito integral $\int G(x_s, s)dW_s$ uses the left endpoint t_k for evaluation. Is this choice unique? Why is the left endpoint chosen, and does it lead to different calculus rules?

Answer:

- **Uniqueness:** No, the choice is not unique. One could choose the midpoint (Stratonovich integral) or the right endpoint.
- **Why Left Endpoint (Ito):** The left endpoint is chosen to preserve the **Martingale property**. This ensures the integral is ”non-anticipating,” meaning the value at time t does not depend on future information. This is crucial in finance and causal systems.

- **Calculus Rules:** Yes, this choice fundamentally changes the rules of calculus. Ito calculus does not follow the standard Chain Rule. Instead, it requires **Ito's Lemma**:

$$df(W_t) = f'(W_t)dW_t + \frac{1}{2}f''(W_t)dt$$

The extra term arises because $(dW_t)^2 = dt$ in the limit.

References:

- Øksendal, B. (2003). *Stochastic Differential Equations: An Introduction with Applications.* (Chapters 3 & 4).

---

3 Question 3: Neural Networks - Odd/Even Functions

Question: In real-world problems, when would we need to enforce that a model must be an odd or even function?

Answer: Enforcing such constraints is a form of **Inductive Bias**.

- **Physics-Informed Machine Learning (PINNs):** When solving Partial Differential Equations (PDEs), if the physical system has symmetry (e.g., a symmetric potential well or heat distribution), the solution must be an even function. Conversely, systems like antisymmetric electromagnetic fields require odd functions.
- **Signal/Image Processing:** In Fourier analysis contexts, real and imaginary parts often correspond to even and odd components.

Enforcing this reduces the search space and improves physical consistency.

References:

- Raissi, M., et al. (2019). *Physics-informed neural networks.*

---

4 Question 4: Logistic Regression - Local Minima

Question: In logistic regression, is it possible for gradient descent to get stuck in a local minimum?

Answer: No.

- Standard logistic regression using the **Cross-Entropy Loss** (Log-Likelihood) is a **Convex Optimization Problem**.

- In convex functions, any local minimum is also the global minimum. Therefore, gradient descent is guaranteed to converge to the global optimum (assuming a proper learning rate), and it will not get stuck in a sub-optimal local pit.

Note: If one were to use Mean Squared Error (MSE) for logistic regression, the landscape would become non-convex, potentially causing local minima issues.

References:

- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* (Chapter 4.3.2).

5 Question 5: Naive Bayes - Diagonal Covariance

Question: "Gaussian Simple Bayesian" assumes a diagonal covariance matrix. Does this mean features are independent? What examples perform better or worse?

Answer:

- **Meaning:** Yes, for Gaussian distributions, a diagonal covariance matrix implies zero correlation, which is equivalent to statistical independence between features ($x_1, \ldots, x_d$).
- **Performs Better:** When features are actually unrelated (e.g., Temperature vs. Zip Code) or when data is scarce (reduces overfitting). It also works surprisingly well on text classification (Bag-of-Words).
- **Performs Worse:** When features are highly correlated (e.g., Height in cm vs. Height in inches). The model will "double count" the evidence, leading to overconfident probability estimates.

References:

- Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective.* (Chapter 3).

6 Question 6: ReLU and Approximation Theory

Question: Classical approximation theory relies on smoothness (Taylor expansions). ReLU is non-differentiable at the origin. Is the theory still applicable?

Answer: Yes, but the mathematical tools differ.

- **Universal Approximation Theorem:** While early proofs (Cybenko, 1989) used sigmoids, Hornik (1991) and Leshno (1993) proved that **any non-polynomial activation function** (including ReLU) can approximate any continuous function arbitrarily well.

- **Explanation:** ReLU networks act as piecewise linear approximations. With enough neurons, these linear segments can fit any curve. The non-differentiability at $x = 0$ is a set of measure zero and does not hinder the integral approximation capabilities. Subgradients are used for optimization at the kink.

References:

- Hornik, K. (1991). *Approximation capabilities of multilayer feedforward networks.*
- Leshno, M., et al. (1993).
