CS109 Lecture 24 Worksheet: ANSWER KEY

Diffusion

Autumn 2026

Lecture 24 Worksheet: Answer Key

Instructor / TA copy.

Problem 1 (Review): What Does the Network Learn?

(a) A sigmoid constrains the classifier's output to $(0, 1)$, appropriate for a Bernoulli probability. A Gaussian noise value can be any real number, including negative values, so imposing that restriction on predicted noise is inappropriate. A neural network need not use a sigmoid at its output; the output should suit the task.

(b) During training, the weights are updated and the training inputs vary across images, noise levels, and fresh noise draws. During generation, the trained weights remain fixed; the current image state and noise level change. The prompt is normally held fixed throughout a run.

(c) **Gradient descent** minimizes a loss: $\theta \leftarrow \theta - \eta \nabla_\theta L$. Gradient ascent maximizes log-likelihood. When the loss is the negative log-likelihood (up to a positive scale and an additive constant), these optimize the same objective in opposite sign conventions. The optional challenge makes the Gaussian squared-error connection explicit.

Problem 2: Make a Noisy Image

(a) $2 \times 2 \times 3 = 12$ entries. x_t is the entire noisy image vector; one coordinate is one pixel's color-channel value.

(b)
$$x_t = 0.8[1, 0] + 0.6[-1, 2] = [0.8, 0] + [-0.6, 1.2] = \boxed{[0.2, 1.2]}.$$
Also, $a_t^2 + b_t^2 = 0.64 + 0.36 = 1$. The weights are not percentages required to sum to 1.

(c) **No.** Gaussian noise has support on the entire real line, so adding it can take coordinates outside the clean-image range. The intermediate noisy state is not required to be a display-ready image.

(d) With $a_t \approx 1, b_t \approx 0$, most original image structure remains. With $a_t \approx 0, b_t \approx 1$, the image is approximately a standard Gaussian noise vector. Different clean inputs approach the same *distribution*, not the same realized array. Independent noise draws still differ.

Teaching note: The constraint controls the scale of a mixture of independent, centered, unit-variance signal and noise: its variance is $a_t^2 + b_t^2 = 1$. A particular fixed image need not have unit variance, and this constraint does not make every individual noisy image have exactly the same empirical variance.

Problem 3: Where Do the Training Labels Come From?

(a) Inputs: the noisy image $[0.2, 1.2]$, the noise level t, and the caption's numerical representation. Target: the sampled noise $\epsilon = [-1, 2]$. The actual sampled noise is used to compute the loss, not supplied as an input. The clean image is used to construct the exercise, not passed as the denoising network's input either.

(b)
$$L = (-1 - (-0.5))^2 + (2 - 1.5)^2 = 0.25 + 0.25 = \boxed{0.5}.$$

(c) No human noise annotation is needed: the program sampled the noise, so it already has the answer key. Fresh noise and different noise levels produce new supervised exercises from the same original. This does not remove the need for suitable images and captions.

(d) The chosen forward process adds noise. The denoising network receives the noisy image, noise level, and caption representation and learns to **predict the noise that was added**. We compare that prediction to the sampled target and update its weights.

Problem 4: An Estimate Is Not a Sample

(a) A is black on the left and white on the right; B is the reverse. Their average is $M = [0.5, 0.5]$, two gray pixels. M is not one of the two possible clean images.

(b) Predicting A gives squared error 0 if A was the original, and
$$(1 - 0)^2 + (0 - 1)^2 = 2$$
if B was the original. Expected error is $\frac{1}{2}(0) + \frac{1}{2}(2) = \boxed{1}$.

(c) Predicting M gives error $0.5^2 + 0.5^2 = 0.5$ for either original. Expected error is therefore $\boxed{0.5}$, smaller than the error for always predicting A (or always predicting B). The conditional mean minimizes expected squared error, even though it need not be a possible clean outcome.

(d) Return A with probability 1/2 and B with probability 1/2. A sample represents one possible outcome; it should not always return M. A fresh randomized guess here also has expected reconstruction error 1, demonstrating that minimizing reconstruction error and sampling the conditional distribution are different goals.

(e) The network supplies an estimate of noise given its inputs. The full sampler uses predictions across many noise levels, starting from random noise, to produce a complete image. At fixed t with $a_t \neq 0$, a predicted noise vector corresponds to a clean-image estimate through $(x_t - b_t\hat{\epsilon})/a_t$. Under ideal squared-error prediction, the conditional mean of the noise corresponds to the conditional mean of the clean image. Returning that single mean is not the same as drawing a sample. A suitable iterative sampling rule can use these estimates to generate different outcomes.

Teaching note: Do not suggest that blindly repeating an average guarantees a valid sample, or that every individual update guarantees better prompt agreement.

Problem 5: One Model, Many Images

(a) Different initial noise arrays can lead to different outcomes. Some samplers also add randomness during intermediate updates. The trained weights do not need to change; ordinary generation does not retrain the network.

(b) This is the model's learned conditional distribution over images given text. Specified details include a frog, a walking pose, stilts supporting it, and a sandy desert. Unspecified details include viewpoint, lighting, the frog's precise appearance, and the arrangement of dunes. Accept any two reasonable examples from each group.

(c) No exact sampled-noise answer key is available for the evolving generated state. Training gave us targets because we deliberately corrupted known clean examples. During generation, a learned noise prediction helps determine an update according to the sampler; it is not guaranteed recovery of a known original image or of the exact initial random draw.

(d) Hold fixed the prompt, model version, sampling procedure, and settings such as image size, number of steps, and guidance strength. A seed initializes pseudorandom draws. Exact reproduction can additionally depend on software, hardware, and implementation details; the seed alone is not a universal image identifier.

Problem 6: How Does the Prompt Become One Scene?

(a) **Frog:** subject appearance and anatomy. **Walking:** a pose suggesting movement. **On stilts:** the frog is supported by stilts, with appropriate contact and placement. **Across a sandy desert:** the setting and the subject's relationship to it. Merely including the objects does not establish that the frog is using the stilts or moving through the scene.

(b) A well-functioning CLIP model should assign greater similarity between the cap's image vector and the matching description than between that image vector and "a blue bicycle." This is an intended learned behavior, not a guarantee for every image. It does not require a dedicated color coordinate: semantic features can be distributed across many coordinates. Matching image and text vectors need not be identical.

(c) Example: "Attention lets the network give different weights to relevant information from the prompt. Subject features can draw on 'frog,' while features of the surroundings draw on 'sandy desert.'" These connections are learned, soft, and overlapping; there is no permanent one-phrase-to-one-region assignment.

(d) The caption influences repeated denoising predictions. The text encoder produces contextual numerical representations of text tokens. The denoising network combines text information with the current noisy image and noise level to predict noise. The sampler uses that prediction to update the image state. CLIP itself does not draw or decode the image, and text embeddings are not simply averaged into four independent object pictures. Cross-attention is a common conditioning mechanism; particular architectures vary.

Problem 7: New Combinations and Imperfect Results

(a) No. Learned patterns can generalize to combinations absent as complete scenes from the training set. Success is not guaranteed: object counts, anatomy, who is eating what, or spatial relationships may be wrong. Generalization does not require an exact training example of the complete request.

(b) The invented word has no agreed visual referent in the prompt. Tokenizing it into familiar pieces does not supply the intended object's appearance. The model may invent something or draw on incidental associations; a description or reference image would make the request clearer.

(c) A visually convincing sample need not obey anatomy or physical constraints. Video additionally requires consistency across frames, including identity, limb count, object placement, and motion.

(d) The image latent is a compressed numerical representation of the visual state that the diffusion process updates. The text representation encodes the prompt and conditions those updates. The decoder converts the final image latent into pixel values. These arrays have different roles and are not interchangeable.

Challenge Problem: Why Squared Error?

(a) Conditional independence gives

$$\ell(\theta) = -\frac{d}{2}\log(2\pi\tau^2) - \frac{1}{2\tau^2}\sum_{j=1}^{d}\left[\epsilon_j - \hat{\epsilon}_{\theta,j}(u)\right]^2.$$

Since τ^2 is fixed and positive, the first term and the multiplier introduce no change to the maximizing weights except the sign: maximizing ℓ is equivalent to minimizing the sum of squared errors. Across training exercises, sum these log-likelihoods. This establishes a Gaussian regression interpretation of the denoising objective, not an exact claim that it equals the complete generated-image log-likelihood.

(b) The variance term is independent of m, and $(\mu - m)^2 \geq 0$, with equality at $m = \mu$. Thus the best fixed prediction is the mean. Applied conditionally to network inputs, the ideal squared-error prediction is the conditional mean of the target, assuming sufficient model capacity and successful optimization.

(c) No. A mean is a single summary; sampling must reflect the distribution's possible outcomes and their probabilities. In Problem 4, $[0.5, 0.5]$ is the best mean estimate, but a sample must be A or B. This is why the denoising estimate and the full sampling procedure play different roles.