Exercise Session 4

DDPMs and Denoising Score Matching

Prepared by Abdulkadir Gokce and Petr Grinberg

Overview

Additional Reading Materials. We recommend following papers:

[1] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "**Denoising Diffusion Probabilistic Models**." Advances in Neural Information Processing Systems (NeurIPS) 33 (2020): 6840-6851.

[2] Song, Yang, et al. "**Score-Based Generative Modeling through Stochastic Differential Equations**." International Conference on Learning Representations (ICLR) (2021).

Task 1. Denoising Diffusion Probabilistic Models (DDPMs).

Diffusion models define a fixed *forward* (noising) Markov chain that gradually destroys structure in the data, and a learned *reverse* (denoising) chain that reconstructs samples. Training proceeds by maximizing a variational lower bound that, under Gaussian assumptions, reduces to a weighted mean-squared denoising objective. This problem develops the key formulas used throughout Ho et al. [1].

1. **Forward marginals remain Gaussian.** We fix a variance schedule $(\beta_t)_{t=1}^T$, set $\alpha_t := 1 - \beta_t$, and $\bar{\alpha}_t := \prod_{s=1}^t \alpha_s$. The forward process is defined as

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}\big(\sqrt{\alpha_t}\,\mathbf{x}_{t-1},\ \beta_t\mathbf{I}\big).$$

   Show that the t-step marginal has the closed form

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}\Big(\sqrt{\bar{\alpha}_t}\,\mathbf{x}_0,\ (1-\bar{\alpha}_t)\mathbf{I}\Big).$$

   Hint: Unroll the recursion and use linearity of Gaussians.

2. **Closed-form forward posterior - *optional*.** Prove that the forward posterior is also Gaussian:

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}\big(\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0),\ \tilde{\beta}_t\mathbf{I}\big), \tag{1}$$

   and derive

$$\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\,\beta_t}{1-\bar{\alpha}_t}\,\mathbf{x}_0 \ + \ \frac{\sqrt{\alpha_t}\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\,\mathbf{x}_t, \qquad \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\,\beta_t.$$

   Hint: Use Bayes: $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \propto q(\mathbf{x}_{t-1}|\mathbf{x}_0)\, q(\mathbf{x}_t|\mathbf{x}_{t-1})$; both are Gaussians in $\mathbf{x}_{t-1}$—multiply them and complete the square to read off $\tilde{\mu}_t$ and $\tilde{\beta}_t$.

3. **Why Gaussian noise?** Briefly explain (2–3 sentences) which parts of subtasks (1)–(2) would fail to admit simple closed forms if the per-step noise were Laplace or uniform instead of Gaussian. What practical consequences would that have for training/sampling?

4. **Variational decomposition (ELBO).** Let the reverse process be

$$p_{\boldsymbol{\theta}}(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\mathbf{x}_{t-1}|\mathbf{x}_t), \quad p(\mathbf{x}_T) = \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Starting from $\log p_{\boldsymbol{\theta}}(\mathbf{x}_0) = \log \int p_{\boldsymbol{\theta}}(\mathbf{x}_{0:T})\, d\mathbf{x}_{1:T}$ and introducing $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$, derive the standard low-variance ELBO split:

$$\begin{aligned} \mathbb{E}_{q(\mathbf{x}_0)}[-\log p_{\boldsymbol{\theta}}(\mathbf{x}_0)] \;\leq\; & \underbrace{\mathbb{E}\, D_{\mathrm{KL}}\big(q(\mathbf{x}_T|\mathbf{x}_0) \,\|\, p(\mathbf{x}_T)\big)}_{\mathcal{L}_T} \\ & + \sum_{t=2}^{T} \underbrace{\mathbb{E}\, D_{\mathrm{KL}}\big(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \,\|\, p_{\boldsymbol{\theta}}(\mathbf{x}_{t-1}|\mathbf{x}_t)\big)}_{\mathcal{L}_{t-1}} \\ & - \underbrace{\mathbb{E} \log p_{\boldsymbol{\theta}}(\mathbf{x}_0|\mathbf{x}_1)}_{\mathcal{L}_0}. \end{aligned} \tag{2}$$

Hint: Insert $\frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}$ to write $\log p_{\boldsymbol{\theta}}(\mathbf{x}_0) = \log \mathbb{E}_q[\cdot]$, apply Jensen inequality to get $\mathbb{E}_q[\log(\cdot)]$, then factor q and use the Gaussian posterior from Eq. 1 to identify the KL terms.

Note: Recall that Kullback–Leibler divergence is defined as $D_{KL}(P\|Q) = \int P(x) \log \frac{P(x)}{Q(x)} dx$.

5. **From mean-matching to ϵ-prediction.** Assume $p_{\boldsymbol{\theta}}(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mu_{\boldsymbol{\theta}}(\mathbf{x}_t, t),\ \sigma_t^2 \mathbf{I})$ with fixed σ_t.

   (a) Show that $\mathcal{L}_{t-1}$ reduces to a weighted MSE between $\mu_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$ and $\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0)$.

   (b) Using the reparameterization $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\, \mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\, \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, prove that the reverse mean can be parameterized as

$$\mu_{\boldsymbol{\theta}}(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\, \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) \right).$$

   (c) Conclude that $\mathcal{L}_{t-1}$ is proportional to $\mathbb{E}\, \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\|_2^2$ with weight $w_t = \frac{\beta_t^2}{2\,\sigma_t^2\,\alpha_t\,(1-\bar{\alpha}_t)}$.

Hint: Substitute the parameterization into the Gaussian KL from Eq. 2.

6. **Simple loss and weighting.** Define the unweighted "simple loss"

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t, \mathbf{x}_0, \boldsymbol{\epsilon}}\, \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\|_2^2, \qquad t \sim \text{Uniform}\{1, \ldots, T\}.$$

Explain in 2–4 sentences why discarding w_t can yield improved sample quality in practice, even if it no longer matches the exact ELBO weighting. Relate your explanation to the SNR profile across t.

7. **Sampling update and $\hat{\mathbf{x}}_0$ estimator.** (a) Derive the ancestral reverse update

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\,\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z}, \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

and list two common choices $\sigma_t \in \{\sqrt{\beta_t},\ \sqrt{\tilde{\beta}_t}\}$. Briefly state a bias–variance trade-off between them.
(b) Show that the "predict-$\mathbf{x}_0$" estimator

$$\hat{\mathbf{x}}_0(\mathbf{x}_t) = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\right)$$

follows directly from the parameterization, and explain its role in diagnostics/progressive reconstructions.

Task 2. From Forward Diffusion to Denoising Score Matching.

This problem derives the *forward diffusion SDE (stochastic differential equation)* from a small-step Gaussian noising process, states the *reverse (generative) SDE* that uses the *score* $\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t)$, and motivates *denoising score matching* as a tractable way to learn that score. The exposition assumes minimal prior exposure to ODE/SDEs and uses first-order (Taylor) reasoning.

Lightweight background (what is an SDE in this context?). A standard ordinary differential equation (ODE) step looks like $\mathbf{x}(t+\Delta t) \approx \mathbf{x}(t) + f(\mathbf{x}(t), t)\,\Delta t$, where f is a drift. An stochastic differential equation (SDE) adds a *random* increment on top:

$$\mathbf{x}(t+\Delta t) \;\approx\; \mathbf{x}(t) \;+\; f(\mathbf{x}(t), t)\,\Delta t \;+\; g(t)\,\sqrt{\Delta t}\,\boldsymbol{\xi}, \quad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Here f (drift) pulls the state in a direction; g (diffusion) sets the noise strength. In the infinitesimal limit, this becomes the SDE $d\mathbf{x}_t = f(\mathbf{x}_t, t)\,dt + g(t)\,d\mathbf{W}_t$, with $\mathbf{W}_t$ a Wiener (Brownian) process.

1. **Small-step forward diffusion ⇒ SDE (via Taylor / first-order limit).**
Consider a fixed *forward* (noising) process with small time step $\Delta t > 0$:

$$\mathbf{x}_{t+\Delta t} \;=\; \sqrt{1-\beta(t)\,\Delta t}\;\mathbf{x}_t \;+\; \sqrt{\beta(t)\,\Delta t}\;\boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

where the schedule $\beta(t) \in (0, 1)$ controls noise injection.

   (a) Use the Taylor expansion $\sqrt{1-\beta(t)\Delta t} \approx 1 - \frac{1}{2}\beta(t)\Delta t$ to show

$$\mathbf{x}_{t+\Delta t} - \mathbf{x}_t \;\approx\; -\tfrac{1}{2}\beta(t)\,\mathbf{x}_t\,\Delta t \;+\; \sqrt{\beta(t)}\,\sqrt{\Delta t}\;\boldsymbol{\epsilon}_t.$$

   (b) Interpret the right-hand side as "deterministic drift $\times\,\Delta t$ + stochastic increment $\times\sqrt{\Delta t}$" and write the forward diffusion *SDE*:

$$d\mathbf{x}_t \;=\; -\tfrac{1}{2}\beta(t)\,\mathbf{x}_t\,dt \;+\; \sqrt{\beta(t)}\,d\mathbf{W}_t\,. \tag{3}$$

   Identify the drift $f(\mathbf{x}, t)$ and diffusion $g(t)$.

Hint: Keep only first-order terms in Δt; the random term scales like $\sqrt{\Delta t}$.

2. **Closed-form forward kernel (why Gaussians are convenient) - *optional*.**
For the SDE in Eq. 3, show that the *conditional* distribution of the noisy sample $\mathbf{x}_t$ given $\mathbf{x}_0$ is Gaussian:

$$q_t(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}\big(\mathbf{x}_t;\ \alpha_t\,\mathbf{x}_0,\ \sigma_t^2\mathbf{I}\big), \quad \alpha_t \ = \ \exp\Big(-\tfrac{1}{2}\int_0^t \beta(s)\,ds\Big), \quad \sigma_t^2 \ = \ 1-\alpha_t^2. \tag{4}$$

Conclude the useful reparameterization

$$\mathbf{x}_t \ = \ \alpha_t\,\mathbf{x}_0 \ + \ \sigma_t\,\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})\,.$$

Why it matters: We can sample x_t directly (one shot), without simulating many tiny steps.

3. **Reverse (generative) SDE — the goal for sampling.**
A classic result [3] gives the *reverse-time* SDE that turns noise back into data:

$$d\mathbf{x}_t \ = \ \Big[-\tfrac{1}{2}\beta(t)\,\mathbf{x}_t \ - \ \beta(t)\,\nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t)\Big]\,dt \ + \ \sqrt{\beta(t)}\,d\overline{\mathbf{W}}_t\,,$$

where time runs backward ($t: T\to 0$) and q_t is the *marginal* density of $\mathbf{x}_t$. Interpret the components of this equation.

4. **Naïve idea: Learn the marginal score directly — why it fails.**
Consider regressing a time-dependent score network $s_\theta(\mathbf{x}_t,t)$ to the *marginal* score:

$$\min_{\boldsymbol{\theta}}\ \mathbb{E}_{t\sim\mathrm{Unif}[0,T]}\ \mathbb{E}_{\mathbf{x}_t\sim q_t}\ \big\|\,s_{\boldsymbol{\theta}}(\mathbf{x}_t,t)-\nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t)\big\|_2^2.$$

Explain briefly why $\nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t)$ is not tractable in general.

Hint: $q_t(\mathbf{x}_t)=\int q_t(\mathbf{x}_t|\mathbf{x}_0)\,q_0(\mathbf{x}_0)\,d\mathbf{x}_0$ is a complicated mixture.

5. **Key trick: Denoising Score Matching (conditional score is tractable).**
Replace the intractable marginal score with the *conditional* score, which is closed-form under Eq. 4:

$$\min_{\theta}\ \mathbb{E}_{t\sim\mathrm{Unif}[0,T]}\ \mathbb{E}_{\mathbf{x}_0\sim q_0}\ \mathbb{E}_{\mathbf{x}_t\sim q_t(\cdot|\mathbf{x}_0)}\big\|\,s_\theta(\mathbf{x}_t,t)-\nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t|\mathbf{x}_0)\big\|_2^2.$$

   (a) Compute the conditional score for the Gaussian kernel:

$$\nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t|\mathbf{x}_0) \ = \ -\frac{1}{\sigma_t^2}\big(\mathbf{x}_t-\alpha_t\mathbf{x}_0\big)\,.$$

   (b) Using the reparameterization $\mathbf{x}_t=\alpha_t x_0+\sigma_t\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})$, simplify to

$$\nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t|x_0) \ = \ -\frac{1}{\sigma_t}\,\boldsymbol{\epsilon}\,.$$

   (c) Conclude the practical *noise-prediction* form: define $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ and set

$$s_{\boldsymbol{\theta}}(\mathbf{x}_t,t) \ = \ -\frac{1}{\sigma_t}\,\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t,t)\,,$$

   so the loss becomes

$$\mathcal{L}(\theta)=\mathbb{E}_{t,x_0,\boldsymbol{\epsilon}}\ \big\|\,\epsilon-\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\alpha_t x_0+\sigma_t\boldsymbol{\epsilon},\ t)\big\|_2^2\,,$$

   optionally with a weighting $\lambda(t)$ in front (to emphasize certain noise levels).

Remark: This is the **same** objective you saw in Task 1 from the DDPM view, now arrived at via the SDE/score perspective!

6. **SDE vs. DDPM perspectives.** In 3–5 sentences, compare when you would prefer the continuous-time SDE/score view over the discrete-time DDPM view, and vice versa. Briefly mention (i) solver flexibility and number of function evaluation, (ii) likelihood/ODE variants (probability flow), and (iii) simplicity of training objectives and schedules in discrete time.

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020.

[2] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations (ICLR)*, 2021.

[3] Brian D.O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. ISSN 0304-4149.