

Here is a concise yet detailed summary of the lecture notes from the PDF, covering the mathematical derivations, theoretical frameworks, and implementation details.

## 1. Score Matching Objectives

The fundamental goal is to train a score network  $S_\theta(x_t, t)$  to approximate the score function  $\nabla_{x_t} \log p_t(x_t)$ . The objective function is the **Fisher Divergence**:

$$\mathcal{L}(\theta) = \int_0^T \lambda(t) \mathbb{E}_{x_t \sim p_t} [||S_\theta(x_t, t) - \nabla_{x_t} \log p_t(x_t)||^2] dt$$

Expanding the squared norm leads to:

$$||S_\theta||^2 - 2\langle S_\theta, \nabla \log p_t \rangle + ||\nabla \log p_t||^2$$

Since the last term does not depend on  $\theta$ , optimizing  $\mathcal{L}(\theta)$  is equivalent to minimizing:

$$J(\theta) = \int_0^T \lambda(t) \mathbb{E}_{x_t} [||S_\theta(x_t, t)||^2 - 2\langle S_\theta(x_t, t), \nabla_{x_t} \log p_t(x_t) \rangle] dt$$

## 2. Implicit Score Matching (ISM) & The Trace Trick

To avoid calculating the unknown  $\nabla \log p_t(x_t)$ , we use integration by parts to rewrite the interaction term.

**Derivation:**

$$\begin{aligned} \mathbb{E}_{x_t} [\langle S_\theta, \nabla \log p_t \rangle] &= \int S_\theta(x)^T \nabla p_t(x) dx = - \int \text{Tr}(\nabla_x S_\theta(x)) p_t(x) dx \\ &= -\mathbb{E}_{x_t} [\nabla \cdot S_\theta(x_t, t)] \end{aligned}$$

where  $\nabla \cdot S_\theta = \text{Tr}(D_x S_\theta(x, t))$  is the divergence (trace of the Jacobian).

The **ISM Loss** becomes:

$$\mathcal{L}_{ISM}(\theta) = \int_0^T \lambda(t) \mathbb{E}_{x_t} [||S_\theta(x_t, t)||^2 + 2\text{Tr}(D_{x_t} S_\theta(x_t, t))] dt$$

## Hutchinson's Trace Estimator

Computing the Jacobian trace is expensive ( $O(d^2)$ ). The notes introduce a stochastic estimator using a random vector  $v$  (e.g.,  $v \sim \mathcal{N}(0, I)$ ):

$$\text{Tr}(A) = \mathbb{E}_v[v^T A v]$$

Substituting  $A = D_{x_t} S_\theta$ :

$$\text{Tr}(D_{x_t} S_\theta) = \mathbb{E}_v[v^T D_{x_t} S_\theta v]$$

## Jacobian-Vector Product (JVP)

Using the definition of the directional derivative,  $v^T D_x S_\theta v$  can be computed efficiently via forward-mode differentiation:

$$v^T D_x S_\theta(x, t)v = v^T \left( \frac{d}{dh} S_\theta(x + hv, t) \Big|_{h=0} \right) = \frac{d}{dh} (v^T S_\theta(x + hv, t)) \Big|_{h=0}$$

### Final ISM Objective:

$$\mathcal{L}(\theta) = \int_0^T \lambda(t) \mathbb{E}_{x_t, v} \left[ \|S_\theta(x_t, t)\|^2 + 2 \frac{d}{dh} v^T S_\theta(x_t + hv, t) \Big|_{h=0} \right] dt$$

## 3. Denoising Score Matching (DSM) & SDEs

The notes transition to diffusion processes defined by a Stochastic Differential Equation (SDE).

### SDE Formulation (Ornstein-Uhlenbeck Process)

$$dX_t = -\beta X_t dt + \sigma dW_t$$

The conditional distribution  $X_t | X_0$  follows a Gaussian distribution:

$$X_t | X_0 \sim \mathcal{N}(\mu_t, \Sigma_t)$$

$$\mu_t = e^{-\beta t} X_0, \quad \Sigma_t = \frac{\sigma^2}{2\beta} (1 - e^{-2\beta t}) I$$

Let  $\gamma_t = e^{-\beta t}$  and  $\sigma_t^2 = \frac{\sigma^2}{2\beta}(1 - e^{-2\beta t})$ . Then:

$$X_t = \gamma_t X_0 + \sigma_t \epsilon, \quad \text{where } \epsilon \sim \mathcal{N}(0, I)$$

## DSM Approximation

DSM replaces the true score  $\nabla \log p_t(x_t)$  with the conditional score  $\nabla \log p_{t|0}(x_t|x_0)$ , which is tractable.

$$\nabla_{x_t} \log p(x_t|x_0) = \nabla_{x_t} \left( -\frac{\|x_t - \gamma_t x_0\|^2}{2\sigma_t^2} \right) = -\frac{x_t - \gamma_t x_0}{\sigma_t^2} = -\frac{\sigma_t \epsilon}{\sigma_t^2} = -\frac{\epsilon}{\sigma_t}$$

## Score Network Parameterization

To stabilize training, we define a "Score Network"  $\epsilon_\theta$  to predict the noise  $\epsilon$ :

$$S_\theta(x_t, t) := -\frac{\epsilon_\theta(x_t, t)}{\sigma_t}$$

Substituting this into the loss function:

$$\mathcal{L}(\theta) = \int_0^T \frac{\lambda(t)}{\sigma_t^2} \mathbb{E}_{x_0, \epsilon} \left[ \left\| -\frac{\epsilon_\theta(\gamma_t x_0 + \sigma_t \epsilon, t)}{\sigma_t} - \left( -\frac{\epsilon}{\sigma_t} \right) \right\|^2 \right] dt$$

$$\mathcal{L}(\theta) = \int_0^T \frac{\lambda(t)}{\sigma_t^2} \mathbb{E}_{x_0, \epsilon} \left[ \frac{1}{\sigma_t^2} \|\epsilon_\theta(x_t, t) - \epsilon\|^2 \right] dt$$

By choosing the weighting function  $\lambda(t) = \sigma_t^2$ , the objective simplifies to pure noise prediction MSE:

$$\mathcal{L}(\theta) = \mathbb{E}_{t, x_0, \epsilon} [\|\epsilon_\theta(\gamma_t x_0 + \sigma_t \epsilon, t) - \epsilon\|^2]$$

## 4. Implementation & Sampling (Reverse Process)

### Training Algorithm

1. Sample time  $t \sim \text{Uniform}(0, T)$ .
2. Sample data  $x_0 \sim p_{\text{data}}$ .
3. Sample noise  $\epsilon \sim \mathcal{N}(0, I)$ .
4. Construct noisy data:  $X_t = \gamma_t x_0 + \sigma_t \epsilon$ .

5. Compute Loss:  $\|\epsilon_\theta(X_t, t) - \epsilon\|^2$ .

6. Update  $\theta$  via Gradient Descent.

## Reverse SDE (Generative Process)

After training  $S_\theta \approx \nabla \log p_t$ , we sample using the reverse-time SDE:

$$d\bar{X}_t = [-\beta\bar{X}_t - \sigma^2 S_\theta(\bar{X}_t, t)]dt + \sigma d\bar{W}_t$$

Substituting the parameterized score  $S_\theta = -\frac{\epsilon_\theta}{\sigma_t}$ :

$$d\bar{X}_t = \left( \frac{\sigma^2}{\sigma_t} \epsilon_\theta(\bar{X}_t, t) - \beta\bar{X}_t \right) dt + \sigma d\bar{W}_t$$

## Discrete Step Update (DDPM-style)

Discretizing the reverse SDE with step size  $\Delta t$ :

$$\bar{X}_{k-1} = \bar{X}_k - \Delta t \left( \frac{\sigma^2}{\sigma_t} \epsilon_\theta(\bar{X}_k, k\Delta t) - \beta\bar{X}_k \right) + \sigma\sqrt{\Delta t}Z_k$$

where  $Z_k \sim \mathcal{N}(0, I)$ . This allows generating samples from noise  $\bar{X}_T \sim \mathcal{N}(0, \sigma_T^2 I)$  back to data  $\bar{X}_0$ .

Here is the concise, detailed summary of the **Denoising Score Matching (DSM)** proof from the lecture notes, with all citations removed as requested.

## Proof of Denoising Score Matching (DSM)

### 1. The Intractable Objective (Explicit Score Matching)

The initial goal is to train a model  $S_\theta(x_t, t)$  to match the true score of the data distribution  $\nabla_{x_t} \log p_t(x_t)$ . The objective function (Fisher Divergence) is:

$$\mathcal{L}(\theta) = \int_0^T \lambda(t) \mathbb{E}_{x_t \sim p_t} [\|S_\theta(x_t, t) - \nabla_{x_t} \log p_t(x_t)\|^2] dt$$

Expanding the squared term, we isolate the interaction term that makes this objective intractable (since  $\nabla \log p_t$  is unknown):

$$\mathcal{L}(\theta) = \int_0^T \lambda(t) \mathbb{E}_{x_t} [\|S_\theta\|^2 - 2\langle S_\theta, \nabla_{x_t} \log p_t(x_t) \rangle + C_1] dt$$

## 2. Key Identity: Marginal vs. Conditional Score

To resolve the intractability, we use the relationship between the marginal score  $u_t(x_t) = \nabla_{x_t} \log p_t(x_t)$  and the conditional score  $\tilde{u}_t(x_t, x_0) = \nabla_{x_t} \log p(x_t|x_0)$ .

The marginal score is the expectation of the conditional score over the posterior  $p(x_0|x_t)$ :

$$\nabla_{x_t} \log p_t(x_t) = \mathbb{E}_{x_0|x_t} [\nabla_{x_t} \log p(x_t|x_0)]$$

## 3. Substitution and Expectation Swap

We substitute this identity into the cross-term of the loss function:

$$\mathbb{E}_{x_t} [\langle S_\theta(x_t), \nabla_{x_t} \log p_t(x_t) \rangle] = \mathbb{E}_{x_t} [\langle S_\theta(x_t), \mathbb{E}_{x_0|x_t} [\nabla_{x_t} \log p(x_t|x_0)] \rangle]$$

Using the law of iterated expectations, we can switch the integration from the marginal  $x_t$  to the joint distribution  $(x_0, x_t)$ :

$$= \mathbb{E}_{x_0, x_t} [\langle S_\theta(x_t), \nabla_{x_t} \log p(x_t|x_0) \rangle]$$

## 4. The Tractable DSM Objective

We plug this term back into the expanded loss function. The objective becomes:

$$\mathcal{L}_{DSM}(\theta) = \int_0^T \lambda(t) \mathbb{E}_{x_0, x_t} [\|S_\theta(x_t, t)\|^2 - 2\langle S_\theta(x_t, t), \nabla_{x_t} \log p(x_t|x_0) \rangle] dt + C$$

By completing the square (adding the constant term  $\|\nabla \log p(x_t|x_0)\|^2$  which is independent of  $\theta$ ), we arrive at the final tractable objective:

$$\mathcal{L}_{DSM}(\theta) = \int_0^T \lambda(t) \mathbb{E}_{x_0} \mathbb{E}_{x_t|x_0} [\|S_\theta(x_t, t) - \nabla_{x_t} \log p(x_t|x_0)\|^2] dt$$

## Conclusion

This derivation proves that minimizing the error with respect to the **conditional score** (which is known and simple, usually a Gaussian kernel) is equivalent to minimizing the error with respect to the true, intractable **data score**.