Appendix A — Proof Sketches of Main Theorems

This appendix offers structured proof sketches for the theoretical results in Section 7. Each theorem outlines key assumptions, logical steps, and references to established results where relevant.

Notation:

- n*: number of training samples;*
- γ*: convergence rate exponent;*
- $\varepsilon_{\mathcal{M}}$*: model class approximation error;*
- β*: KL weight in VAEs;*
- T*: diffusion steps;*
- L*: Lipschitz constant of score network.*
- δ: Confidence level (used in probabilistic bounds)
- α: Model-specific convergence constant
- p_{data}: True data distribution
- p_G: Model-generated distribution
- $\mathcal{D}$: Divergence metric (e.g., KL, JS, Wasserstein)
- E: Expectation operator
- $ELBO$: Evidence Lower Bound
- FID: Fréchet Inception Distance
- MCR: Mode Coverage Ratio
- $\mathcal{R}_n$: Rademacher complexity on n samples
- F_G, F_D: Generator and discriminator function classes
- $s(x), s'(x)$: Score functions learned from perturbed datasets

Theorem 7.1 (Unified Convergence Rate)

Let $M \in$ {VAE, GAN, Diffusion} denotes a generative model trained on n samples drawn from the data distribution p_{data}. Then, under standard regularity assumptions:

$$E\left[D\left(p_{\text{data}},\, p_{\theta^{(n)}}\right)\right] \leq C_{\mathcal{M}} \cdot n^{-\alpha_{\mathcal{M}}} + \varepsilon_{\mathcal{M}} \quad \text{(A.1)}$$

Where:

- $C_{\mathcal{M}}$ is a model-specific constant
- $\alpha_{\mathcal{M}}$ is the convergence exponent
- $\varepsilon_{\mathcal{M}}$ is the model class approximation error

We observe the ordering:

$$\alpha_{\text{Diffusion}} \geq \alpha_{\text{VAE}} \geq \alpha_{\text{GAN}} \quad \text{(A.2)}$$

Sketch: This follows from uniform convergence bounds and classical VC-type inequalities, assuming bounded loss and Lipschitz continuity of model families. The diffusion case invokes SDE regularity; the GAN case applies minimax generalization bounds.

Theorem 7.2 (Lower Bound on Sample Quality)

$$\text{Quality}(p_\theta) \geq \text{GIE}(p_\theta, p_{\text{data}}) - \mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right) \ \text{(A.3)}$$

Sketch: The proof exploits the convexity of the ELBO and the KL-divergence penalty. Under over-regularization (large β), posterior collapse occurs, which geometrically excludes modes from the latent space. Taylor expansion of the KL term bounds the deviation.

Theorem 7.3 (Sample Complexity Bound):

To achieve distributional approximation error ε with confidence 1 − δ, the required number of samples satisfies: (See Appendix A for proof sketch.)

$$n(\varepsilon, \delta) = O\left(\frac{C(H) \cdot d_{\text{eff}} \cdot \log(1/\delta)}{\varepsilon^2}\right) \quad \text{(A.4)}$$

Sketch: Using optimal transport theory and gradient penalties, the Wasserstein distance ensures smooth alignment between $p_d ata$ and p_G. Collapse corresponds to Jacobian singularity. Bounding the discriminator's curvature stabilizes mode inclusion.

Theorem 7.4 (Adversarial Generalization Bound):

With probability at least $1 - \delta$:

$$\left| D_{\text{JS}}(p_{\text{data}}, p_G) - \widehat{D_{\text{JS}}^{(n)}} \right| \leq \mathcal{O}\left(\frac{\mathcal{R}_n(G \circ D) + \log(1/\delta)}{n}\right) \qquad \text{(A.5)}$$

Where $\mathcal{R}_n(G \circ D)$ is the empirical Rademacher complexity of the composition of the generator and discriminator networks.

Sketch: Based on the score matching objective and the SDE formulation, diffusion models approximate the target distribution via repeated noisy refinements. A coupling argument over time steps yields a concentration result around all modes, with a convergence rate

~ $1/\sqrt{T}$.

Theorem 7.5 (Diffusion Stability):

Let s_θ and $s_{\theta'}$ be score functions learned on datasets that differ in k samples. Then:

$$|s_\theta - s_{\theta'}|^2 \leq \left(\frac{n}{2k}\right) \cdot L \cdot \sqrt{T} \quad \text{(A.6)}$$

Where L is the Lipschitz constant of the score network, and T is the diffusion length. This proves that training stability improves as the sample size increases

Sketch: The result follows from the first-order expansion of the divergence function (e.g., KL or JS) under perturbations, assuming differentiability and strong convexity. The generator inherits robustness through parameter continuity.