KL Comparison with Forward Process and Pushforward Weighting

Mark Ogata

October 2025

1 Aligned Parameter-Gradient Forms with Forward Process and Noise Averaging

We compare the parameter derivatives of the **forward KL**, **reverse KL**, and our **pushforward score** objective under a forward diffusion process $F(x,t)$ and noise-level expectation $\mathbb{E}_t$ (which is exactly the procedure we use to compute DMD loss in practice).

Forward KL

$$\mathbb{E}_t\Big[D_{\mathrm{KL}}(p_{\mathrm{fake}}\|p_{\mathrm{real}})\Big] = \mathbb{E}_t\left[\int p_{\mathrm{fake}}(x)\big(\log p_{\mathrm{fake}}(x) - \log p_{\mathrm{real}}(x)\big)\,dx\right]$$

$$\Rightarrow \quad \frac{d}{d\theta}\mathbb{E}_t\Big[D_{\mathrm{KL}}(p_{\mathrm{fake}}\|p_{\mathrm{real}})\Big] = \mathbb{E}_t\left[\int \underbrace{p_{\mathrm{fake}}(x)}_{\text{marginal}}\underbrace{\big(s_{\mathrm{fake}}(F(x,t),t) - s_{\mathrm{real}}(F(x,t),t)\big)}_{\text{score diff}}\frac{dG_\theta(x_t)}{d\theta}\,dx\right].$$

Reverse KL

$$\mathbb{E}_t\Big[D_{\mathrm{KL}}(p_{\mathrm{real}}\|p_{\mathrm{fake}})\Big] = \mathbb{E}_t\left[\int p_{\mathrm{real}}(x)\big(\log p_{\mathrm{real}}(x) - \log p_{\mathrm{fake}}(x)\big)\,dx\right]$$

$$\Rightarrow \quad \frac{d}{d\theta}\mathbb{E}_t\Big[D_{\mathrm{KL}}(p_{\mathrm{real}}\|p_{\mathrm{fake}})\Big] = \mathbb{E}_t\left[\int \underbrace{p_{\mathrm{real}}(x)}_{\text{marginal}}\underbrace{\big(s_{\mathrm{real}}(F(x,t),t) - s_{\mathrm{fake}}(F(x,t),t)\big)}_{\text{score diff}}\frac{dG_\theta(x_t)}{d\theta}\,dx\right].$$

Our Objective

$$\frac{d}{d\theta}L_{\mathrm{ours}} = -\,\mathbb{E}_t\left[\int \underbrace{r(x)}_{\text{weight}}\underbrace{\big(s_{\mathrm{real}}(F(x,t),t) - s_{\mathrm{fake}}(F(x,t),t)\big)}_{\text{score diff}}\frac{dG_\theta(x_t)}{d\theta}\,dx\right].$$

Comparison Table

	Marginal / Weight	Score difference	Outer expectation
Forward KL	$p_{\mathrm{fake}}(x)$	$(s_{\mathrm{fake}} - s_{\mathrm{real}})$	$\mathbb{E}_t$
Reverse KL	$p_{\mathrm{real}}(x)$	$(s_{\mathrm{real}} - s_{\mathrm{fake}})$	$\mathbb{E}_t$
Ours	$r(x)$	$(s_{\mathrm{real}} - s_{\mathrm{fake}})$	$\mathbb{E}_t$

Thus, our objective shares the same **score difference** pattern as the reverse KL, but is evaluated under a different **marginal weighting** $r(x)$.

2 Why the Pushforward r is Not a Linear Combination of p_{data} or p_θ

We recall the definition of the pushforward weighting:

$$r(x) = \int p_\theta(x \mid x_t)\, q(x_t \mid x_0)\, p_{\text{data}}(x_0)\, dx_t\, dx_0.$$

It mixes samples by first drawing $x_0 \sim p_{\text{data}}$, then diffusing to $x_t \sim q(x_t \mid x_0)$, and finally partially denoising via $p_\theta(x \mid x_t)$.

We ask: could this mixture be written as a simple convex combination

$$r(x) \stackrel{?}{=} \alpha\, p_{\text{data}}(x) + (1-\alpha)\, p_\theta(x), \quad \text{for some } \alpha \in [0,1]?$$

Step 1: What r really is

The integrand shows that r is a **mixture over many conditionals** $p_\theta(x \mid x_t)$, each weighted by how likely x_t is under diffused data $q(x_t \mid x_0)p_{\text{data}}(x_0)$. Explicitly,

$$r(x) = \int w(x_t)\, p_\theta(x \mid x_t)\, dx_t, \quad \text{where} \quad w(x_t) = \int q(x_t \mid x_0)\, p_{\text{data}}(x_0)\, dx_0.$$

Thus, r is an average of *different shapes of the model conditional*, each corresponding to a different noise level or latent x_t.

Step 2: Why this is not a linear combination of marginals

The key difference is that a convex combination of p_{data} and p_θ,

$$\alpha\, p_{\text{data}}(x) + (1-\alpha)\, p_\theta(x),$$

is a mixture of only *two fixed distributions*, each with a fixed shape in x.

In contrast, $r(x)$ integrates infinitely many *varying* distributions $p_\theta(x \mid x_t)$, whose means, variances, or even supports change with x_t. Formally, unless all $p_\theta(x \mid x_t)$ share the same functional form as $p_\theta(x)$, the mixture

$$r(x) = \int w(x_t)\, p_\theta(x \mid x_t)\, dx_t$$

cannot be represented as a single affine combination of two fixed distributions in x.

Step 3: Simple example (Gaussian case)

Suppose

$$q(x_t \mid x_0) = \mathcal{N}(x_t; x_0, \sigma_t^2 I), \quad p_\theta(x \mid x_t) = \mathcal{N}(x; \mu_\theta(x_t), \Sigma_t).$$

Then $r(x)$ is a Gaussian mixture:

$$r(x) = \int w(x_t)\, \mathcal{N}(x; \mu_\theta(x_t), \Sigma_t)\, dx_t.$$

Unless $\mu_\theta(x_t)$ is constant (so all components are identical), the result is a *mixture of Gaussians with different means*, not a single Gaussian shape. Therefore, r cannot equal any convex combination of two fixed densities p_{data} and p_θ—its shape has higher-order variability.

Step 4: When equality can hold

Equality could only occur in degenerate situations:

- If $p_\theta(x \mid x_t) = p_\theta(x)$ for all x_t, then $r(x) = p_\theta(x)$.
- If $p_\theta(x \mid x_t) = \delta(x - x_0)$ (perfect reconstruction), then $r(x) = p_{\text{data}}(x)$.
- If $p_{\text{data}} = p_\theta$, then all distributions coincide.

Conclusion

In general,

$$r(x) \;\neq\; \alpha\, p_{\text{data}}(x) + (1-\alpha)\, p_\theta(x),$$

because r is not built from two fixed shapes—it is the *pushforward mixture* of p_{data} through the joint stochastic map $(x_0 \to x_t \to x)$, producing a distribution with structure that lies strictly between p_{data} and p_θ, but not on their linear span. Since integration is a linear operation, this means we cannot write our loss as a linear combination of the forward and backward KL divergences.