Drifting Models

Yutao Chen

⧖ May 19 2026
⧗ May 19 2026

Contents

Drifting models (Deng et al., 2026) are generative models that admits fast one-step sampling by minimizing *drifting* at train time.

Drifting Field

Let $p_{\text{data}}(\boldsymbol{x})$ be the data distribution, and $q_{\boldsymbol{\theta}}(\boldsymbol{x})$ be the model distribution induced by a neural network $\boldsymbol{x} = f_{\boldsymbol{\theta}}(\boldsymbol{z})$ with noise $\boldsymbol{z} \sim p(\boldsymbol{z})$. Drifting models approximate $p(\boldsymbol{x})$ with $q_{\boldsymbol{\theta}}(\boldsymbol{x})$ by minimizing

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}\sim q_{\boldsymbol{\theta}}}\left[\|\boldsymbol{x} - \text{sg}\big(\boldsymbol{x} + V_{p,q}(\boldsymbol{x})\big)\|^2\right],$$

where $\text{sg}(\cdot)$ is the stop-gradient operator, and $V_{p,q}(\boldsymbol{x})$ is the drifting field.

The drifting field $V_{p,q}(\boldsymbol{x})$ for distributions p, q on $\mathbb{R}^d$ is defined as

$$V_{p,q}(\boldsymbol{x}) = V_p(\boldsymbol{x}) - V_q(\boldsymbol{x}),$$

with

$$V_p(\boldsymbol{x}) = \frac{\mathbb{E}_{\boldsymbol{y}\sim p}[k(\boldsymbol{x},\boldsymbol{y})\boldsymbol{y}]}{\mathbb{E}_{\boldsymbol{y}\sim p}[k(\boldsymbol{x},\boldsymbol{y})]} - \boldsymbol{x}, \quad V_q(\boldsymbol{x}) = \frac{\mathbb{E}_{\boldsymbol{y}\sim q}[k(\boldsymbol{x},\boldsymbol{y})\boldsymbol{y}]}{\mathbb{E}_{\boldsymbol{y}\sim q}[k(\boldsymbol{x},\boldsymbol{y})]} - \boldsymbol{x},$$

where $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a kernel function. Intuitively, $V_p(\boldsymbol{x})$ / $V_q(\boldsymbol{x})$ can be interpreted as *attraction / repulsion* forces respectively, where the former pulls $q_{\boldsymbol{\theta}}(\boldsymbol{x})$ towards $p(\boldsymbol{x})$ and the latter disperses $q_{\boldsymbol{\theta}}$.

In practice, we approximate $V_{p,q}(\boldsymbol{x})$ empirically by drawing batches of attracting samples $\{\boldsymbol{y}^+\} \sim p$ and repulsing samples $\{\boldsymbol{y}^-\} \sim q$.

Theoretical Connections

Score Matching

The drifting field $V_{p,q}(\boldsymbol{x})$ can be interpreted as an approximation of the score mismatch for densities p, q with *Gaussian* kernels.

Specifically, consider Gaussian kernels with temperature $\tau > 0$

$$k(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{\left(\sqrt{2\pi}\tau\right)^d} \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|_2^2}{2\tau^2}\right).$$

The kernel smoothed (convolved) density of $p(\boldsymbol{x})$ on $\mathbb{R}^d$ is defined as

$$\tilde{p}(\boldsymbol{x}) \coloneqq (p * k)(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{y}\sim p}[k(\boldsymbol{x}, \boldsymbol{y})].$$

For Gaussian kernels k, we can show that $V_{p,q}(\boldsymbol{x})$ is proportional to the score mismatch between the kernel smoothed densities $\tilde{p}, \tilde{q}$

$$V_{p,q}(\boldsymbol{x}) \propto \nabla_{\boldsymbol{x}} \log \tilde{p}(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} \log \tilde{q}(\boldsymbol{x}).$$

Proof. Note that $\nabla_{\boldsymbol{x}} k(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{\tau^2} k(\boldsymbol{x}, \boldsymbol{y})(\boldsymbol{y} - \boldsymbol{x})$, and that

$$\nabla_{\boldsymbol{x}} \log \tilde{p}(\boldsymbol{x}) = \frac{\nabla_{\boldsymbol{x}} \tilde{p}(\boldsymbol{x})}{\tilde{p}(\boldsymbol{x})}, \qquad \nabla_{\boldsymbol{x}} \tilde{p}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{y}\sim p}[\nabla_{\boldsymbol{x}} k(\boldsymbol{x}, \boldsymbol{y})].$$

Therefore, we have

$$\nabla_{\boldsymbol{x}} \log \tilde{p}(\boldsymbol{x}) = \frac{1}{\tau^2} \frac{\mathbb{E}_{\boldsymbol{y}\sim p}[k(\boldsymbol{x}, \boldsymbol{y})(\boldsymbol{y} - \boldsymbol{x})]}{\mathbb{E}_{\boldsymbol{y}\sim p}[k(\boldsymbol{x}, \boldsymbol{y})]} = \frac{1}{\tau^2} V_p(\boldsymbol{x}).$$

Similarly, we can show that $\nabla_{\boldsymbol{x}} \log \tilde{q}(\boldsymbol{x}) = \frac{1}{\tau^2} V_q(\boldsymbol{x})$. Therefore, we have

$$V_{p,q}(\boldsymbol{x}) = V_p(\boldsymbol{x}) - V_q(\boldsymbol{x}) = \tau^2 (\nabla_{\boldsymbol{x}} \log \tilde{p}(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} \log \tilde{q}(\boldsymbol{x})).$$

Reverse KL Divergence

Consider the that the Wasserstein gradient flow of the reverse KL divergence $\mathcal{F}(q) = \mathbb{D}_{\mathrm{KL}}(q \parallel p)$. The functional derivative w.r.t. $q(\boldsymbol{x})$ is

$$\frac{\partial \mathcal{F}}{\partial q}(\boldsymbol{x}) = 1 + \log q(\boldsymbol{x}) - \log p(\boldsymbol{x}),$$

and the corresponding gradient field w.r.t. $\boldsymbol{x}$ is given by

$$\nabla_{\boldsymbol{x}} \frac{\partial \mathcal{F}}{\partial q}(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} \log q(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}).$$

We can see the drifting field $V_{p,q}(\boldsymbol{x})$ yields the steepest *descent* directon of $\mathbb{D}_{\mathrm{KL}}(\tilde{q} \parallel \tilde{p})$ for kernel smoothed densities $\tilde{p}, \tilde{q}$.

> When assuming kernels are *characteristic*, minimizing $\mathbb{D}_{\mathrm{KL}}(\tilde{q} \parallel \tilde{p})$ is equivalent to minimizing $\mathbb{D}_{\mathrm{KL}}(q \parallel p)$.

We refer readers to Gretton et al. (2026) for further discussions.

Kernel-Gradient Drifting

Note that the connection between the drifting field and score mismatch

$$V_{p,q}(\boldsymbol{x}) \approx \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} \log q(\boldsymbol{x})$$

only holds for Gaussian kernels, which satisfy $\nabla_{\boldsymbol{x}} k(\boldsymbol{x}, \boldsymbol{y}) \propto (\boldsymbol{y} - \boldsymbol{x})$.

For general kernels, we consider replacing $(\boldsymbol{y} - \boldsymbol{x})$ with $\nabla_{\boldsymbol{x}} \log k(\boldsymbol{x}, \boldsymbol{y})$, yielding the kernel-gradient drifting in Esteban-Casadevall et al. (2026):

$$V_{p,q}(\boldsymbol{x}) = \frac{\mathbb{E}_{\boldsymbol{y}\sim p}[\nabla_{\boldsymbol{x}} k(\boldsymbol{x}, \boldsymbol{y})]}{\mathbb{E}_{\boldsymbol{y}\sim p}[k(\boldsymbol{x}, \boldsymbol{y})]} - \frac{\mathbb{E}_{\boldsymbol{y}\sim q}[\nabla_{\boldsymbol{x}} k(\boldsymbol{x}, \boldsymbol{y})]}{\mathbb{E}_{\boldsymbol{y}\sim q}[k(\boldsymbol{x}, \boldsymbol{y})]},$$

noting that $k(\boldsymbol{x}, \boldsymbol{y}) \nabla_{\boldsymbol{x}} \log k(\boldsymbol{x}, \boldsymbol{y}) = \nabla_{\boldsymbol{x}} k(\boldsymbol{x}, \boldsymbol{y})$.

References

Deng, M., Li, H., Li, T., Du, Y., & He, K. (2026,). *Generative Modeling via Drifting*. https://arxiv.org/abs/2602.04770

Esteban-Casadevall, M., Carrasco-Pollo, J., Welling, M., Meent, J.-W. van de, Bekkers, E. J., & Eijkelboom, F. (2026,). *Kernel-Gradient Drifting Models*. https://arxiv.org/abs/2605.10727

Gretton, A., Wenliang, L. K., Galashov, A., Thornton, J., Bortoli, V. D., & Doucet, A. (2026,). *On the Wasserstein Gradient Flow Interpretation of Drifting Models*. https://arxiv.org/abs/2605.05118