Improved sampling via learned diffusions

Lorenz Richter*, Julius Berner*

ICLR

May, 2024

Generative modeling and sampling

Task

Sample from a complex (high-dimensional, multimodal) distribution p_{target}.

Generative modeling and sampling

Task

Sample from a complex (high-dimensional, multimodal) distribution p_{target}.

p_{target} can be given in the form of:

1. **samples** $X^{(i)} \sim p_{\text{target}}$ (images, video, audio, text, ...).

Generative modeling and sampling

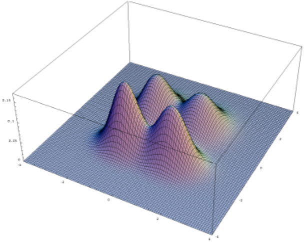
Task

Sample from a complex (high-dimensional, multimodal) distribution p_{target}.

p_{target} can be given in the form of:

1. **samples** $X^{(i)} \sim p_{\text{target}}$ (images, video, audio, text, ...).

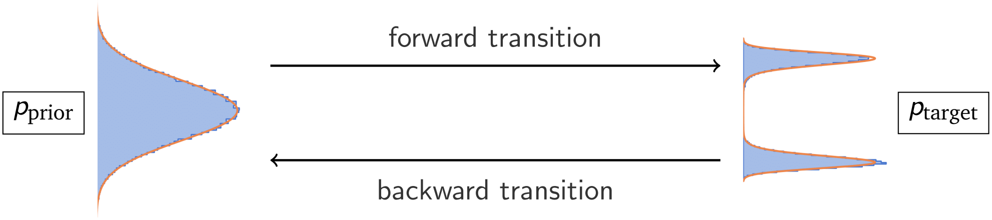
2. an (unnormalized) **density** $p_{\text{target}} = \rho/\mathcal{Z}$ (e.g., in Bayesian statistics, computational physics and chemistry).



Goal: Sample from a distribution $p_{\text{target}} = \rho/\mathcal{Z}$ using an auxiliary distribution p_{prior}.

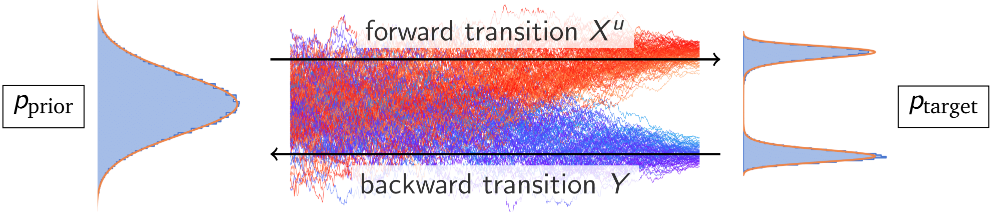
Sampling via learned diffusions

Goal: Sample from a distribution $p_{\text{target}} = \rho/\mathcal{Z}$ using an auxiliary distribution p_{prior}.



Sampling via learned diffusions

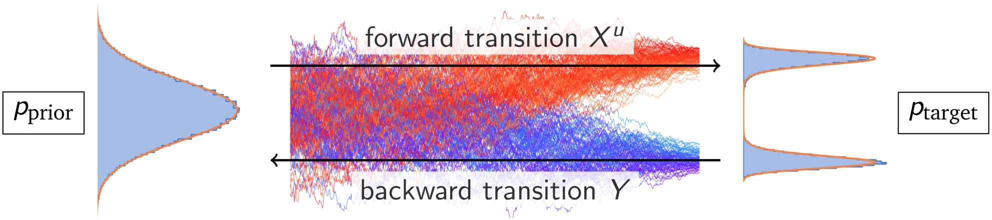
Goal: Sample from a distribution $p_{\mathrm{target}} = \rho/\mathcal{Z}$ using an auxiliary distribution p_{prior}.



Setting: Consider controlled SDEs (with the notation $\overleftarrow{\sigma}(t) := \sigma(T-t)$)

$$\mathrm{d}X_s^u = (\mu + \sigma u)(X_s^u, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s, \qquad X_0^u \sim p_{\mathrm{prior}},$$

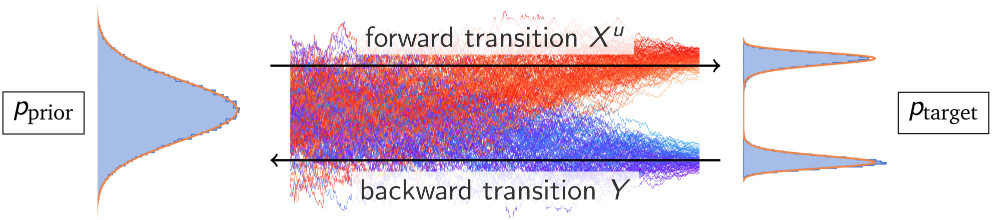
Goal: Sample from a distribution $p_{\text{target}} = \rho/\mathcal{Z}$ using an auxiliary distribution p_{prior}.



Setting: Consider controlled SDEs (with the notation $\overleftarrow{\sigma}(t) \coloneqq \sigma(T-t)$)

$$\mathrm{d}X_s^u = (\mu + \sigma u)(X_s^u, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s, \qquad X_0^u \sim p_{\text{prior}},$$
$$\mathrm{d}Y_s = -\overleftarrow{\mu}(Y_s, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}W_s, \qquad Y_0 \sim p_{\text{target}}.$$

Goal: Sample from a distribution $p_{\mathrm{target}} = \rho / \mathcal{Z}$ using an auxiliary distribution p_{prior}.



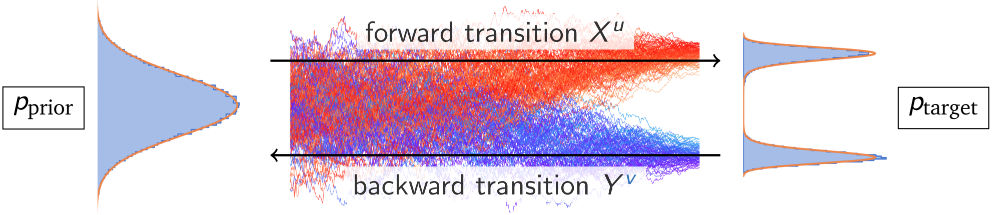
Setting: Consider controlled SDEs (with the notation $\overleftarrow{\sigma}(t) := \sigma(T-t)$)

$$
\begin{aligned}
\mathrm{d}X_s^u &= (\mu + \sigma u)(X_s^u, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s, \qquad & X_0^u &\sim p_{\mathrm{prior}},\\
\mathrm{d}Y_s &= -\overleftarrow{\mu}(Y_s, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}W_s, \qquad & Y_0 &\sim p_{\mathrm{target}}.
\end{aligned}
$$

Idea: Learn u s.t. X^u is the time-reversal of Y, implying $X_T^u \sim p_{\mathrm{target}}$ if $Y_T \sim p_{\mathrm{prior}}$.

Sampling via learned diffusions

Goal: Sample from a distribution $p_{\text{target}} = \rho/\mathcal{Z}$ using an auxiliary distribution p_{prior}.



Setting: Consider controlled SDEs (with the notation $\overleftarrow{\sigma}(t) := \sigma(T-t)$)

$$\mathrm{d}X_s^u = (\mu + \sigma u)(X_s^u, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s, \qquad X_0^u \sim p_{\text{prior}},$$
$$\mathrm{d}Y_s^v = (-\overleftarrow{\mu} + \overleftarrow{\sigma}\overleftarrow{v})(Y_s^v, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}W_s, \qquad Y_0^v \sim p_{\text{target}}.$$

Idea: Learn u, v s.t. X^u is the time-reversal of Y^v, implying $X_T^u \sim p_{\text{target}}$ and $Y_T^v \sim p_{\text{prior}}$.

Sampling via learned diffusions

- Two controlled SDEs:

$$
\begin{aligned}
\mathrm{d}X_s^u &= (\mu + \sigma u)(X_s^u, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s, & X_0^u &\sim p_{\mathrm{prior}}, \\
\mathrm{d}Y_s^v &= (-\overleftarrow{\mu} + \overleftarrow{\sigma}\overleftarrow{v})(Y_s^v, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}W_s, & Y_0^v &\sim p_{\mathrm{target}}.
\end{aligned}
$$

Sampling via learned diffusions

- Two controlled SDEs:

$$\mathrm{d}X_s^u = (\mu + \sigma u)(X_s^u, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s, \qquad X_0^u \sim p_{\mathrm{prior}},$$
$$\mathrm{d}Y_s^v = (-\overleftarrow{\mu} + \overleftarrow{\sigma}\overleftarrow{v})(Y_s^v, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}W_s, \qquad Y_0^v \sim p_{\mathrm{target}}.$$

- Considering the **general case** ($v \neq 0$) one can:
  - take (arbitrary) informed priors,

Sampling via learned diffusions

- Two controlled SDEs:

$$\mathrm{d}X_s^u = (\mu + \sigma u)(X_s^u, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s, \qquad X_0^u \sim p_{\mathrm{prior}},$$
$$\mathrm{d}Y_s^v = (-\overleftarrow{\mu} + \overleftarrow{\sigma}\overleftarrow{v})(Y_s^v, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}W_s, \qquad Y_0^v \sim p_{\mathrm{target}}.$$

- Considering the **general case** ($v \neq 0$) one can:
  - take (arbitrary) informed priors,
  - remove prior errors since $Y_T \sim p_{\mathrm{prior}}$ only holds approximately,

Sampling via learned diffusions

- Two controlled SDEs:

$$\mathrm{d}X_s^u = (\mu + \sigma u)(X_s^u, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s, \qquad X_0^u \sim p_{\mathrm{prior}},$$
$$\mathrm{d}Y_s^v = (-\overleftarrow{\mu} + \overleftarrow{\sigma}\overleftarrow{v})(Y_s^v, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}W_s, \qquad Y_0^v \sim p_{\mathrm{target}}.$$

- Considering the **general case** ($v \neq 0$) one can:
  - take (arbitrary) informed priors,
  - remove prior errors since $Y_T \sim p_{\mathrm{prior}}$ only holds approximately,
  - use (arbitrary) informed coefficients σ, μ and time horizons T.

Sampling via learned diffusions

- Two controlled SDEs:

$$
\begin{aligned}
\mathrm{d}X_s^u &= (\mu + \sigma u)(X_s^u, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s, \qquad && X_0^u \sim p_{\mathrm{prior}}, \\
\mathrm{d}Y_s^v &= (-\overleftarrow{\mu} + \overleftarrow{\sigma}\overleftarrow{v})(Y_s^v, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}W_s, \qquad && Y_0^v \sim p_{\mathrm{target}}.
\end{aligned}
$$

- Considering the **general case** ($v \neq 0$) one can:
  - take (arbitrary) informed priors,
  - remove prior errors since $Y_T \sim p_{\mathrm{prior}}$ only holds approximately,
  - use (arbitrary) informed coefficients σ, μ and time horizons T.
- When only given an **unnormalized density** ρ, we **cannot use**:
  - score matching or variants of likelihood training (no samples from p_{target}),

Sampling via learned diffusions

- Two controlled SDEs:

$$\mathrm{d}X_s^u = (\mu + \sigma u)(X_s^u, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s, \qquad X_0^u \sim p_{\mathrm{prior}},$$
$$\mathrm{d}Y_s^v = (-\overleftarrow{\mu} + \overleftarrow{\sigma}\overleftarrow{v})(Y_s^v, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}W_s, \qquad Y_0^v \sim p_{\mathrm{target}}.$$

- Considering the **general case** ($v \neq 0$) one can:
  - take (arbitrary) informed priors,
  - remove prior errors since $Y_T \sim p_{\mathrm{prior}}$ only holds approximately,
  - use (arbitrary) informed coefficients σ, μ and time horizons T.
- When only given an **unnormalized density** ρ, we **cannot use**:
  - score matching or variants of likelihood training (no samples from p_{target}),
  - reverse KL between density of X_T^u and p_{target} as in continuous normalizing flows (probability flow ODE is only accurate at convergence).

Sampling via learned diffusions

- Two controlled SDEs:

$$\mathrm{d}X_s^u = (\mu + \sigma u)(X_s^u, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s, \qquad X_0^u \sim p_{\mathrm{prior}},$$
$$\mathrm{d}Y_s^v = (-\overleftarrow{\mu} + \overleftarrow{\sigma}\overleftarrow{v})(Y_s^v, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}W_s, \qquad Y_0^v \sim p_{\mathrm{target}}.$$

- Considering the **general case** ($v \neq 0$) one can:
  - take (arbitrary) informed priors,
  - remove prior errors since $Y_T \sim p_{\mathrm{prior}}$ only holds approximately,
  - use (arbitrary) informed coefficients σ, μ and time horizons T.
- When only given an **unnormalized density** ρ, we **cannot use**:
  - score matching or variants of likelihood training (no samples from p_{target}),
  - reverse KL between density of X_T^u and p_{target} as in continuous normalizing flows (probability flow ODE is only accurate at convergence).

Idea: Use divergences between **path measures**.

Sampling via learned diffusions: Path measures

- Two controlled SDEs:

$$\mathrm{d}X_s^u = (\mu + \sigma u)(X_s^u, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s, \qquad X_0^u \sim p_{\mathrm{prior}},$$
$$\mathrm{d}Y_s^v = (-\overleftarrow{\mu} + \overleftarrow{\sigma}\overleftarrow{v})(Y_s^v, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}W_s, \qquad Y_0^v \sim p_{\mathrm{target}}.$$

- **Path space perspective:** Consider path measures $\mathbb{P}_{X^u}$ and $\mathbb{P}_{\overleftarrow{Y}^v}$ on $C([0,T],\mathbb{R}^d)$.

Sampling via learned diffusions: Path measures

- Two controlled SDEs:

$$\mathrm{d}X_s^u = (\mu + \sigma u)(X_s^u, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s, \qquad X_0^u \sim p_{\mathrm{prior}},$$
$$\mathrm{d}Y_s^v = (-\overleftarrow{\mu} + \overleftarrow{\sigma}\overleftarrow{v})(Y_s^v, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}W_s, \qquad Y_0^v \sim p_{\mathrm{target}}.$$

- **Path space perspective:** Consider path measures $\mathbb{P}_{X^u}$ and $\mathbb{P}_{\overleftarrow{Y}^v}$ on $C([0,T],\mathbb{R}^d)$.
- Identify controls u^*, v^* via **divergence** D between those measures

$$u^*, v^* \in \operatorname*{arg\,min}_{u,v} D\big(\mathbb{P}_{X^u} \big| \mathbb{P}_{\overleftarrow{Y}^v}\big).$$

Sampling via learned diffusions: Path measures

- Two controlled SDEs:

$$\mathrm{d}X_s^u = (\mu + \sigma u)(X_s^u, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s, \qquad X_0^u \sim p_{\mathrm{prior}},$$
$$\mathrm{d}Y_s^v = (-\overleftarrow{\mu} + \overleftarrow{\sigma}\overleftarrow{v})(Y_s^v, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}W_s, \qquad Y_0^v \sim p_{\mathrm{target}}.$$

- **Path space perspective:** Consider path measures $\mathbb{P}_{X^u}$ and $\mathbb{P}_{\overleftarrow{Y}^v}$ on $C([0,T],\mathbb{R}^d)$.
- Identify controls u^*, v^* via **divergence** D between those measures

$$u^*, v^* \in \operatorname*{arg\,min}_{u,v} D\big(\mathbb{P}_{X^u} \big| \mathbb{P}_{\overleftarrow{Y}^v}\big).$$

Proposition (Log-likelihood for path measures)

$$\log \frac{\mathrm{d}\mathbb{P}_{X^u}}{\mathrm{d}\mathbb{P}_{\overleftarrow{Y}^v}}(X^w) = \int_0^T \Big((u+v)\cdot\Big(w + \frac{v-u}{2}\Big) + \nabla\cdot(\sigma v - \mu)\Big)(X_s^w, s)\,\mathrm{d}s$$
$$+ \int_0^T (u+v)(X_s^w, s)\cdot \mathrm{d}W_s + \log \frac{p_{\mathrm{prior}}(X_0^w)}{p_{\mathrm{target}}(X_T^w)}$$

- Note that solutions $u^*, v^* \in \arg\min_{u,v} D\left(\mathbb{P}_{X^u} \middle| \mathbb{P}_{\overleftarrow{Y}^v}\right)$ are **not unique**.

Sampling via learned diffusions: Special cases

- Note that solutions $u^*, v^* \in \arg\min_{u,v} D\left(\mathbb{P}_{X^u} \middle| \mathbb{P}_{\overleftarrow{Y}^v}\right)$ are **not unique**.
- We can make them unique by, e.g.,

Sampling via learned diffusions: Special cases

- Note that solutions $u^*, v^* \in \arg\min_{u,v} D\left(\mathbb{P}_{X^u} \middle| \mathbb{P}_{\overleftarrow{Y}^v}\right)$ are **not unique**.
- We can make them unique by, e.g.,
  - fixing the control $v = 0$ (DIS),

Sampling via learned diffusions: Special cases

- Note that solutions $u^*, v^* \in \arg\min_{u,v} D\big(\mathbb{P}_{X^u} \big| \mathbb{P}_{\overleftarrow{Y}^v}\big)$ are **not unique**.
- We can make them unique by, e.g.,
  - fixing the control $v = 0$ (DIS),
  - using time-reversals of suitable reference processes (PIS, DDS),

Sampling via learned diffusions: Special cases

- Note that solutions $u^*, v^* \in \arg\min_{u,v} D\big(\mathbb{P}_{X^u}\big|\mathbb{P}_{\overleftarrow{Y}^v}\big)$ are **not unique**.
- We can make them unique by, e.g.,
  - fixing the control $v = 0$ (DIS),
  - using time-reversals of suitable reference processes (PIS, DDS),
  - Schrödinger bridge: Adding regularizer, e.g., $D_{\mathrm{KL}}(\mathbb{P}_{X^u}|\mathbb{P}_{X^0}) = \mathbb{E}\left[\frac{1}{2}\int_0^T \|u(X_s^u, s)\|^2 \,\mathrm{d}s\right]$.

Sampling via learned diffusions: Special cases

- Note that solutions $u^*, v^* \in \arg\min_{u,v} D\left(\mathbb{P}_{X^u} \middle| \mathbb{P}_{\overleftarrow{Y}^v}\right)$ are **not unique**.
- We can make them unique by, e.g.,
  - fixing the control $v = 0$ (DIS),
  - using time-reversals of suitable reference processes (PIS, DDS),
  - Schrödinger bridge: Adding regularizer, e.g., $D_{\mathrm{KL}}(\mathbb{P}_{X^u}|\mathbb{P}_{X^0}) = \mathbb{E}\left[\frac{1}{2}\int_0^T \|u(X_s^u, s)\|^2 \,\mathrm{d}s\right]$.
- Popular choice for divergence D:

$$D_{\mathrm{KL}}(\mathbb{P}_{X^u}|\mathbb{P}_{\overleftarrow{Y}^v}) = \mathbb{E}\left[\log \frac{\mathrm{d}\mathbb{P}_{X^u}}{\mathrm{d}\mathbb{P}_{\overleftarrow{Y}^v}}(X^u)\right].$$

Sampling via learned diffusions: Special cases

- Note that solutions $u^*, v^* \in \arg\min_{u,v} D\left(\mathbb{P}_{X^u} \middle| \mathbb{P}_{\overleftarrow{Y}^v}\right)$ are **not unique**.
- We can make them unique by, e.g.,
  - fixing the control $v = 0$ (DIS),
  - using time-reversals of suitable reference processes (PIS, DDS),
  - Schrödinger bridge: Adding regularizer, e.g., $D_{\mathrm{KL}}(\mathbb{P}_{X^u}|\mathbb{P}_{X^0}) = \mathbb{E}\left[\frac{1}{2}\int_0^T \|u(X_s^u, s)\|^2 \,\mathrm{d}s\right]$.
- Popular choice for divergence D:

$$D_{\mathrm{KL}}(\mathbb{P}_{X^u}|\mathbb{P}_{\overleftarrow{Y}^v}) = \mathbb{E}\left[\log \frac{\mathrm{d}\mathbb{P}_{X^u}}{\mathrm{d}\mathbb{P}_{\overleftarrow{Y}^v}}(X^u)\right].$$

- **Problems:**
  - Requires to simulate X^u and differentiate through the SDE solver (on-policy).

Sampling via learned diffusions: Special cases

- Note that solutions $u^*, v^* \in \arg\min_{u,v} D\left(\mathbb{P}_{X^u} \middle| \mathbb{P}_{\overleftarrow{Y}^v}\right)$ are **not unique**.
- We can make them unique by, e.g.,
  - fixing the control $v = 0$ (DIS),
  - using time-reversals of suitable reference processes (PIS, DDS),
  - Schrödinger bridge: Adding regularizer, e.g., $D_{\mathrm{KL}}(\mathbb{P}_{X^u}|\mathbb{P}_{X^0}) = \mathbb{E}\left[\frac{1}{2}\int_0^T \|u(X_s^u, s)\|^2 \,\mathrm{d}s\right]$.
- Popular choice for divergence D:

$$D_{\mathrm{KL}}(\mathbb{P}_{X^u}|\mathbb{P}_{\overleftarrow{Y}^v}) = \mathbb{E}\left[\log \frac{\mathrm{d}\mathbb{P}_{X^u}}{\mathrm{d}\mathbb{P}_{\overleftarrow{Y}^v}}(X^u)\right].$$

- **Problems:**
  - Requires to simulate X^u and differentiate through the SDE solver (on-policy).
  - Known to suffer from mode collapse.

Path space perspective: Different divergences

- We propose a novel divergence:

- We propose a novel divergence:

Definition (Log-variance divergence)

$$D_{\mathrm{LV}}^{w}(\mathbb{P}_{X^u}, \mathbb{P}_{\overleftarrow{Y}^v}) := \mathrm{Var}\left(\log \frac{\mathrm{d}\mathbb{P}_{X^u}}{\mathrm{d}\mathbb{P}_{\overleftarrow{Y}^v}}(X^w)\right)$$

Path space perspective: Different divergences

- We propose a novel divergence:

Definition (Log-variance divergence)

$$D_{\mathrm{LV}}^{w}(\mathbb{P}_{X^u}, \mathbb{P}_{\overleftarrow{Y}^v}) := \operatorname{Var}\left(\log \frac{\mathrm{d}\mathbb{P}_{X^u}}{\mathrm{d}\mathbb{P}_{\overleftarrow{Y}^v}}(X^w)\right)$$

- **Off-policy:** In principle arbitrary choice for w allows to balance exploration and exploitation.

Path space perspective: Different divergences

- We propose a novel divergence:

Definition (Log-variance divergence)

$$D_{\mathrm{LV}}^{w}(\mathbb{P}_{X^u}, \mathbb{P}_{\overleftarrow{Y}^v}) := \mathrm{Var}\left(\log \frac{\mathrm{d}\mathbb{P}_{X^u}}{\mathrm{d}\mathbb{P}_{\overleftarrow{Y}^v}}(X^w)\right)$$

- **Off-policy:** In principle arbitrary choice for w allows to balance exploration and exploitation.
- **Zero-th order:** more efficient since no differentiation through the SDE solver and no gradients of p_{target} are necessary.

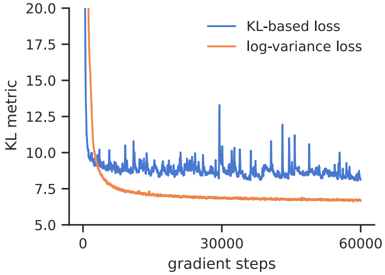
Path space perspective: Different divergences

- We propose a novel divergence:

Definition (Log-variance divergence)

$$D_{\mathrm{LV}}^{w}(\mathbb{P}_{X^u}, \mathbb{P}_{\overleftarrow{Y}^v}) := \mathrm{Var}\left(\log \frac{\mathrm{d}\mathbb{P}_{X^u}}{\mathrm{d}\mathbb{P}_{\overleftarrow{Y}^v}}(X^w)\right)$$

- **Off-policy:** In principle arbitrary choice for w allows to balance exploration and exploitation.
- **Zero-th order:** more efficient since no differentiation through the SDE solver and no gradients of p_{target} are necessary.
- **Sticking-the-landing:** Variance reduction due to control variate property.



Gaussian mixture

Better mode coverage: Improved performance with D_{LV} (compared against D_{KL}) for PIS, DIS, DDS, and the general bridge sampler.

Gaussian mixture

Better mode coverage: Improved performance with D_{LV} (compared against D_{KL}) for PIS, DIS, DDS, and the general bridge sampler.

Method	Divergence	$\Delta \log \mathcal{Z}(rw) \downarrow$	$\mathcal{W}_\gamma^2 \downarrow$	ESS $\uparrow$	Δstd $\downarrow$
CRAFT		0.012	0.020	-	0.019
PIS	KL	0.249	0.467	0.0051	1.937
	LV	**0.001**	**0.020**	**0.9093**	**0.023**
DIS	KL	**0.015**	0.064	0.0226	2.522
	LV	0.017	**0.020**	**0.8660**	**0.004**
DDS	KL	0.005	0.042	0.0737	2.161
	LV	**0.001**	**0.020**	**0.8929**	**0.006**
Bridge	KL	0.560	0.393	0.0180	0.698
	LV	**0.100**	**0.020**	**0.9669**	**0.010**

Better mode coverage: Improved performance with D_{LV} (compared against D_{KL}) for PIS, DIS, DDS, and the general bridge sampler.

Method	Divergence	$\Delta \log \mathcal{Z}(rw) \downarrow$	$\mathcal{W}_\gamma^2 \downarrow$	ESS $\uparrow$	Δstd $\downarrow$
CRAFT		0.012	<u>0.020</u>	-	0.019
PIS	KL	0.249	0.467	0.0051	1.937
	LV	**<u>0.001</u>**	**<u>0.020</u>**	**0.9093**	**0.023**
DIS	KL	**0.015**	0.064	0.0226	2.522
	LV	0.017	**<u>0.020</u>**	**0.8660**	**<u>0.004</u>**
DDS	KL	0.005	0.042	0.0737	2.161
	LV	**<u>0.001</u>**	**<u>0.020</u>**	**0.8929**	**0.006**
Bridge	KL	0.560	0.393	0.0180	0.698
	LV	**0.100**	**<u>0.020</u>**	**<u>0.9669</u>**	**0.010**

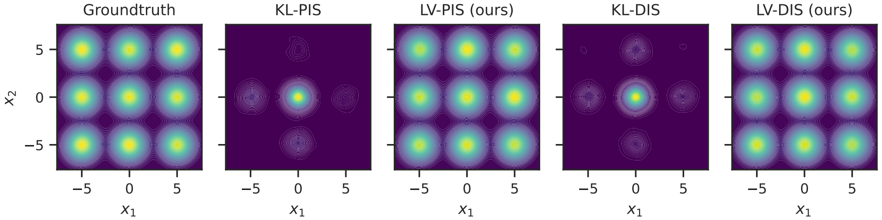


Image density and funnel distribution

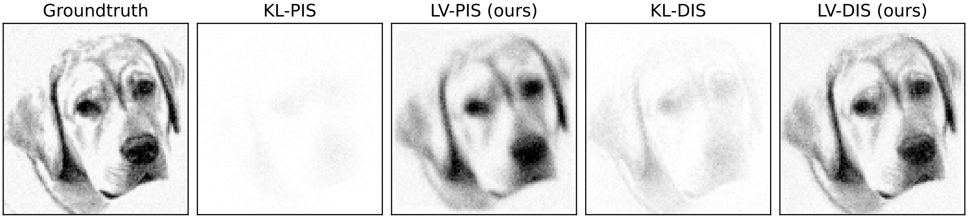
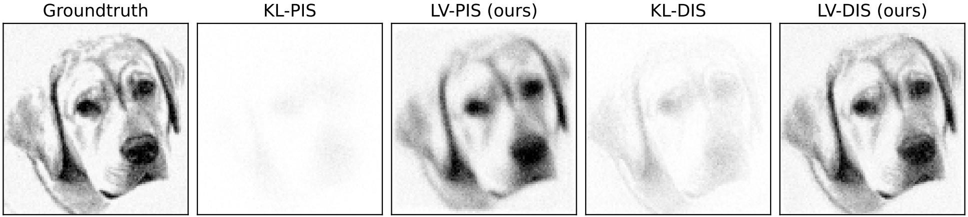


Image density and funnel distribution



Funnel distribution (challenging benchmark for sampling methods):

$$p_{\mathrm{target}}(x) = \mathcal{N}(x_1; 0, 9) \prod_{i=2}^{10} \mathcal{N}(x_i; 0, e^{x_1})$$

Image density and funnel distribution

Groundtruth | KL-PIS | LV-PIS (ours) | KL-DIS | LV-DIS (ours)

Funnel distribution (challenging benchmark for sampling methods):

$$p_{\text{target}}(x) = \mathcal{N}(x_1; 0, 9) \prod_{i=2}^{10} \mathcal{N}(x_i; 0, e^{x_1})$$

Method	Divergence	$\Delta \log \mathcal{Z}(rw) \downarrow$	$\mathcal{W}_\gamma^2 \downarrow$	ESS $\uparrow$	Δstd $\downarrow$
CRAFT		0.123	5.517	-	6.139
PIS	KL	0.111	5.639	**0.1333**	6.921
	LV	**0.097**	**5.593**	0.0746	**6.852**
DIS	KL	0.032	5.120	0.1383	5.254
	LV	**0.028**	**5.075**	**0.2313**	**5.224**
DDS	KL	0.045	5.305	0.1446	6.133
	LV	**0.043**	5.305	**0.1999**	**6.123**

Image density and funnel distribution

Funnel distribution (challenging benchmark for sampling methods):

$$p_{\text{target}}(x) = \mathcal{N}(x_1; 0, 9) \prod_{i=2}^{10} \mathcal{N}(x_i; 0, e^{x_1})$$

Method	Divergence	$\Delta \log \mathcal{Z}(rw) \downarrow$	$\mathcal{W}_\gamma^2 \downarrow$	ESS $\uparrow$	Δstd $\downarrow$
CRAFT		0.123	5.517	-	6.139
PIS	KL	0.111	5.639	**0.1333**	6.921
	LV	**0.097**	**5.593**	0.0746	**6.852**
DIS	KL	0.032	5.120	0.1383	5.254
	LV	**0.028**	**5.075**	**0.2313**	**5.224**
DDS	KL	0.045	5.305	0.1446	6.133
	LV	**0.043**	5.305	**0.1999**	**6.123**

Our improved samplers based on diffusion processes are competitive with state-of-the-art methods based on SMC & normalizing flows (CRAFT).

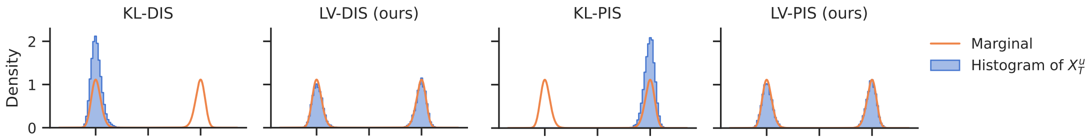
Many-Well problem

Many-Well: typical problem in molecular dynamics with

$$\log \rho(x) = -\sum_{i=1}^{w}(x_i^2 - \delta)^2 - \frac{1}{2}\sum_{i=w+1}^{d} x_i^2.$$

Many-Well: typical problem in molecular dynamics with

$$\log \rho(x) = -\sum_{i=1}^{w}(x_i^2 - \delta)^2 - \frac{1}{2}\sum_{i=w+1}^{d} x_i^2.$$

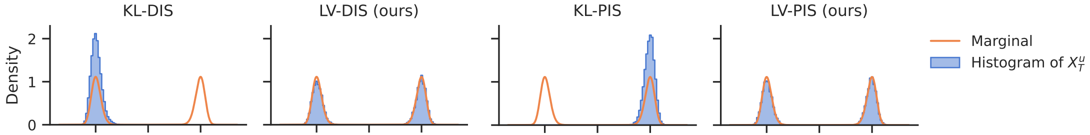


Many-Well problem

Many-Well: typical problem in molecular dynamics with

$$\log \rho(x) = -\sum_{i=1}^{w}(x_i^2 - \delta)^2 - \frac{1}{2}\sum_{i=w+1}^{d} x_i^2.$$

Problem	Method	$\Delta \log Z \downarrow$	$\mathcal{W}_\gamma^2 \downarrow$	ESS $\uparrow$	Δ std $\downarrow$
Many-Well	PIS-KL	3.567	1.699	0.0004	1.409
($d = 5, w = 5, \delta = 4$)	**PIS-LV**	**0.214**	**0.121**	**0.6744**	**0.001**
	DIS-KL	1.462	1.175	0.0012	0.431
	DIS-LV	**0.375**	**0.120**	**0.4519**	**0.001**
Many-Well	PIS-KL	0.101	**6.821**	0.8172	0.001
($d = 50, w = 5, \delta = 2$)	**PIS-LV**	**0.087**	6.823	**0.8453**	**0.000**
	DIS-KL	1.785	**6.854**	0.0225	0.009
	DIS-LV	**1.783**	6.855	**0.0227**	0.009



Thank you for your attention!

Github: `https://github.com/juliusberner/sde_sampler`
Mail: `richter@zib.de`, `mail@jberner.info`

References:

- J. Berner, L. Richter, K. Ullrich. *An optimal control perspective on diffusion-based generative modeling*. TMLR, 2024.
- L. Richter., J. Berner. *Improved sampling via learned diffusions*. ICLR, 2024.