Supplementary Material for “Variational Annealing of GANs: A Langevin Perspective”

C Tao[1], S Dai[1], L Chen[1], K Bai[1], J Chen[3], C Liu[2], R Zhang[1], G Bobashev[4], L Carin[1]
[1]Duke, [2]Tsinghua, [3]Fudan & [4]RTI

A Derivation for Eqn (4-5)

$$
\begin{aligned}
F_\mu(\rho;\beta) &= \mathrm{KL}(\rho \parallel \mu) + (1-\beta)\mathbb{E}_\rho[\log \mu] && (1)\\
&= \beta \mathrm{KL}(\rho \parallel \mu) && (2)\\
&\qquad + (1-\beta)\mathbb{E}_\rho[\log \rho - \log \mu] && (3)\\
&\qquad + (1-\beta)\mathbb{E}_\rho[\log \mu] && (4)\\
&= \beta \mathrm{KL}(\rho \parallel \mu) + (1-\beta)\mathbb{E}_\rho[\log \rho] && (5)\\
&= \beta \mathrm{KL}(\rho \parallel \mu) + (1-\beta)S(\rho) && (6)
\end{aligned}
$$

B Proof of Corollary 2.2

Proof. We only need to $\beta_{\text{lik}} = 1+\lambda$ and $\beta_{\text{ent}} = 1/(1+\lambda)$ respectively for the two types of regularizations. For the likelihood regularization, based on (4) we have

$$\beta_{\text{lik}}(1-\beta_{\text{lik}}^{-1}) = \lambda \Rightarrow \beta_{\text{lik}} = \lambda + 1,$$

and similarly for entropy regularization we have

$$\beta_{\text{ent}}^{-1} - 1 = \lambda \Rightarrow \beta_{\text{ent}} = 1/(1+\lambda)$$

from (5). This concludes our proof. □

C Continuity Equation

In the following, we omit the dependency on space-time to avoid notational clutter. By the divergence theorem, a general continuity equation takes the following differential form:

$$\partial_t \rho + \nabla \cdot \boldsymbol{j} = \sigma, \tag{7}$$

where ρ is the quantity of substance per unit volume, and $\boldsymbol{j}$ is the flux of the substance and σ is the generation/absorption rate per unit volume per unit time. When $\boldsymbol{v}$ is a velocity field the describes the flow $\boldsymbol{j}$, then $\boldsymbol{j} = \rho \boldsymbol{v}$. Since the mass of probability distribution is conserved, we have $\sigma(x,t) \equiv 0$. Pluging these terms into the continuity equation gives us

$$\partial_t \rho + \nabla \cdot (\rho \boldsymbol{v}) = 0. \tag{8}$$

D Proof of Theorem 2.3

Proof. Notice

$$
\begin{aligned}
&\mathrm{KL}(\rho_t \parallel \mu) && (9)\\
=\ &\mathrm{KL}(\rho_t \parallel \mu_\beta) + \mathbb{E}_{X' \sim \rho_t}[\log \mu_\beta(X') - \log \mu(X')] &&\\
\leq\ &\mathrm{KL}(\rho_t \parallel \mu_\beta) + \delta_\beta. && (10)
\end{aligned}
$$

To get the $\mathcal{O}(t^{-1})$ convergence rate for first term, simply apply Theorem 4.5 from Liu (2017). This result can be improved to exponential decay assuming stronger regularity conditions on μ. For instance, when $\mathrm{KL}(\cdot||\mu)$ is geodesically strongly convex (e.g., when the density of μ is strongly log-concave on $\mathcal{X}$ (Villani (2008), Theorem 17.15)) on the 2-Wasserstein space $\mathcal{P}_2$ (Villani (2008), Definition 6.4), the gradient flow ρ_t of $\mathrm{KL}(\cdot||\mu)$ will converge exponentially (Villani (2008), Theorem 23.25 & 24.7). □

E Further Notes on RKL-GAN as gradient flow

It is helpful to further the understanding of the dynamics from a gradient flow perspective. Consider minimizing $F(q) : \mathcal{P} \to \mathbb{R}$ be a function on the space of probability measure $\mathcal{P}(\mathcal{X})$, in our case, the anneal RKL. At each step, we would like to find a proper perturbation $\partial_t q_t$ for updating q. Formally, $\partial_t q_t$ should be an element in the tangent space of $\mathcal{P}(\mathcal{X})$ at q.

To practically describe $\partial_t q_t$, people notice that a $\partial_t q_t$ is related to a bunch of velocity fields $\{v_t\}$ on $\mathcal{X}$ through the continuity equation (6), and a one-to-one relation can be established by choosing a particular v_t with the minimal norm in the bunch, when a proper norm is defined for velocity fields on $\mathcal{X}$ (e.g., when the norm is taken as the one of $L_q^2(\mathcal{X};\mathbb{R}^n)$, the one-to-one relation (unique existance of v_t) is guaranteed by e.g. Villani (2008), Theorem 13.8; Ambrosio et al. (2008), Theorem 8.3.1, Proposition 8.4.5.) Also note that the description with v_t automatically satisfies the restriction on $\partial_t q_t$: $\int_{\mathcal{X}} \partial_t q_t \, \mathrm{d}x = -\int_{\mathcal{X}} \nabla \cdot (q_t v_t) \, \mathrm{d}x = -\int_{\partial \mathcal{X}} q_t v_t \cdot \mathrm{d}\overrightarrow{S} = 0$, where $\overrightarrow{S}$ is the infinitesimal directed surface area on the boundary $\partial \mathcal{X}$.

With this description, we can write the directional derivative of the RKL F with respect to $\partial_t q_t$, or v_t: $\frac{\mathrm{d}}{\mathrm{d}s}F(q+s\partial_t q_t)|_{s=0} = \mathbb{E}_q[-\nabla\cdot v_t + \Psi\cdot v_t]$, as is shown in Theorem 3.1 of Liu and Wang (2016).

There is more that the velocity field description could provide. Let $\mathcal{T}$ be the set of all representative v_t's, which is a linear space. With a proper inner product so that $\mathcal{T}$ is a Hilbert space, $T_q\mathcal{P}(\mathcal{X})$ will be a Hilbert space due to the one-to-one relationship (which is now an isometric isomorphism), which (roughly) means that $\mathcal{P}(\mathcal{X})$ is a Riemannian manifold. With the Riemannian structure, gradient of F on $\mathcal{P}(\mathcal{X})$ can be defined, which is characterized by $\frac{\mathrm{d}}{\mathrm{d}s}F(q+s\partial_t q_t)|_{s=0} = \langle \operatorname{grad} F, \partial_t q_t\rangle_{T_q\mathcal{P}(\mathcal{X})}$, or

$$\operatorname{grad} F = \max\cdot \underset{\|\partial_t q_t\|_{T_q\mathcal{P}(\mathcal{X})}=1}{\arg\max}\left\{\frac{\mathrm{d}}{\mathrm{d}s}F(q+s\partial_t q_t)|_{s=0}\right\}. \tag{11}$$

When $\mathcal{T}$ is taken as $L^2_q(\mathcal{X};\mathbb{R}^n)$ and $\mathcal{P}(\mathcal{X})$ as 2-Wasserstein space, we have $\operatorname{grad} F(q) = -\Psi - \nabla\log q$ (Villani (2008), Theorem 23.18; Ambrosio et al. (2008), Example 11.1.2), i.e. the tangent vector on the gradient flow of $F(q)$. Note that the Langevin dynamics (1) is also known as the dynamics of the gradient flow of $F(q)$ on 2-Wasserstein space Jordan et al. (1998), since it produces the same $\partial_t q_t$ through the Fokker-Planck equation as the dynamics of $\operatorname{grad} F$ through the continuity equation (6). Since the $\log q_t$ term here is intractable, it is estimated in (annealed) RKL-GAN through a function approximator (*e.g.* neural net) updated by stochastic gradient descent (the critic updates).

The quality of the approximation affects the empirical convergence rate, as the asymptotic convergence rate is an upper bound on the improvement, and it only holds if the updates are exactly aligned with functional gradients. So the maximizing step can be understood as searching for the best descent direction for the generator update. Another choice of $\mathcal{T}$ is $\mathcal{H}^n$ where $\mathcal{H}$ is the RKHS of a kernel k, as is done in Liu and Wang (2016); Liu (2017). *(Note that in this case, $\mathcal{T}\subset T_q\mathcal{P}(\mathcal{X})$, which means that the existence of v_t in $\mathcal{T}$ is not guaranteed for any $\partial_t q_t \in T_q\mathcal{P}(\mathcal{X})$! Moreover, $\mathcal{P}(\mathcal{X})$ as a set is not defined in this case!)* The gradient in this case is then $\operatorname{grad} F(q) = \mathbb{E}_q(x)[-\Psi(x)k(x,\cdot) + \nabla_x k(x,\cdot)]$, which is the tangent vector to the gradient flow of $F(q)$ on $\mathcal{P}(\mathcal{X})$ defined in Liu (2017). With the help of a kernel, the gradient here is tractable.

F Score Function Estimator (Fig. 2)

We compare different score function estimators with toy examples. We train a DAE with samples from the *banana distribution* $\log u(x_1,x_2) = -\frac{1}{2}(x_2^2/4 + (x_1 - x_2^2/4)^2)$. We set the noise level to $\sigma = 0.2$. In Figure 2 from main text, we compare the model samples drawn from the standard Langevin system using the DAE-score estimator $s_\sigma(x) = \sigma^{-2}(\phi_\sigma(x) - x)$ and the DAE-residual estimator $s^r_\sigma(x) = -\nabla\|\phi_\sigma(x) - x\|^2_2$. The residual estimator is unable to faithfully recover the underlying distribution, for practical training setups, often collapsing samples on modes that not even necessarily align with the real data modes.

G More on generative flows

Typically, the shift-scale flow are given as

$$z_{k+1} = t(z_k) + s(z_k)\odot z_k, \tag{12}$$

where $t(z), s(z)$ respects the causal dependency of an auto-regressive flow, *i.e.*, $[f(z)]_d = f_d([z]_{<d})$. This ensures Jacobians are triangular by design. Equation (12) is a so-called *forward flow* which is fast in generation but slow in evaluation, because the backward process (*e.g.*, $x\to z$) resembles a Gaussian elimination process. Such trade-offs are common in more expressive generative flows. MAF instead exploits the *backward flow* described in the main text.

While it is tempting to train a flow ν directly towards μ, there is one caveat. To optimize $\mathrm{KL}(\nu\parallel\mu)$, one must sample from ν, evaluate the log-likelihood $\log\nu(x)$ and then back-propagate the gradient through it. As discussed above, computational efficiency of sampling and evaluation for popular choices of generative flows are generally at odds with each other. Coupling the two will always invoke the back-propagation of the more costly part. On the other hand, the scheme we have developed avoids back-propagating the more challenging part via amortizing it through a free-form generator.

H Additional Comments

While the U-GAN can be much more flexible than the paired generative flow, we found that more expressive flow allows much quicker learning for U-GAN.

I Experimental Setups

I.1 Variational annealing

In this experiment, we use the DFM implementation from `https://github.com/pfnet-research/chainer-gan-lib` as our codebase. The original model corresponds to the JSD-GAN results reported, and we changed the objective function to derive the RKL-GAN. For the W-GAN experiments, we adapted

Table 1: Network architectures used for Cifar10 experiments. BN: batch normalization, lReLU: Leaky ReLU.

Denoiser	Generator	Discriminator
Input feature f	Input z	Input X
MLP output 2048, lReLU, BN	MLP output 2048, ReLU, BN	3×3 conv. 32 lReLU, stride 1, BN
MLP output 2048, lReLU, BN		4×4 conv. 64 lReLU, stride 2, BN
MLP output 2048, lReLU, BN	4×4 deconv. 256 ReLU, stride 2, BN	4×4 conv. 128 lReLU, stride 2, BN
MLP output 2048, lReLU, BN	4×4 deconv. 128 ReLU, stride 2, BN	4×4 conv. 256 lReLU, stride 2, BN
MLP output 2048, lReLU, BN	4×4 deconv. 64 ReLU, stride 2, BN	4×4 conv. 512 lReLU, stride 2, BN
MLP output 2048, lReLU, BN	3×3 deconv. 3 ReLU, stride 1, BN	Output feature f with shape $512 \times 2 \times 2$
MLP output 2048, Reshape to $512 \times 2 \times 2$		MLP output 1

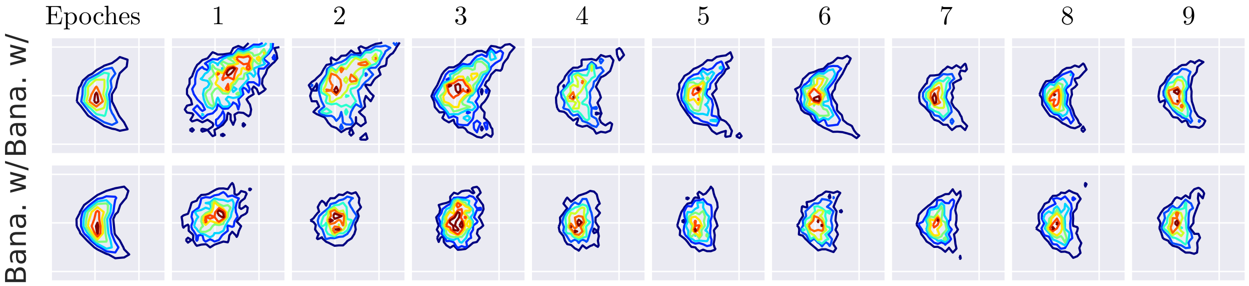

Figure 1: Learning from an unnormalized density to sample the banana distribution.

the code of the WGAN-GP implementation from the same repository, which uses gradient penalty to enforce the Lipschitz constraint. We have used the same hyper-parameter settings from the DFM repository, which potentially explained the fact that the JSD-GAN delivered the best performance in our experiments, as the released DFM code was fine-tuned for this objective. The Adam optimizer with a learning rate of 10^{-4} was used, and all models are trained for about 200 epochs with batch-size 64. Detailed model architecture for Cifar10 (32×32) is summarized in Table 1. Similar architectures are used for CelebA (64×64) and high-res CelebA (128×128), with additional deconvolutional layers.

Dynamic annealing schemes In the dynamic annealing experiments, we found that the linear annealing

$$\lambda_t = \mathrm{sign}(\lambda_0) \max\{|\lambda_0|(1 - t/T_{\max}), 0\}$$

worked better than the exponential annealing

$$\lambda_t = \mathrm{sign}(\lambda_0)|\lambda_0| \exp(-\tau t)$$

among monotonic schemes, where $T_{\max}$ and τ are hyper-parameters controlling the descent rate. $T_{\max}$ was set to be slightly smaller than the total number of iterations to finalize the training, and τ was set chosen so that at end of training $\lambda_t \to 0$. As for the oscillatory annealing, we evaluated stepwise (discontinuous) and sinuous (continuous) designs, and the later worked much better. The performance also depends on $|\lambda_0|$, in our experiments, we have tested $|\lambda_0| \in \{0.1, 1, 10\}$ and reported the best result. During the entire training, λ_t loops over for five cycles.

Table 2: Results for Bayesian Logistic regression.

Dataset	Features	Train	Test
Cancer	32	285	284
Heart	13	135	135
German	20	500	500
Sonar	60	104	104

I.2 Unnormalized GAN

I.2.1 Toy distribution

The *kidney distribution* is given by

$$\log u_{\mathrm{kid}}(x) = \frac{1}{2} s_1(x)^2 - \log\left(s_2(x) + s_3(x)\right),$$

where $s_1(x) = \frac{\|x\|-2}{0.4}$, $s_2(x) = \exp(-\frac{1}{2}[\frac{x_1-2}{0.6}]^2)$ and $s_3(x) = \exp(-\frac{1}{2}[\frac{x_1+2}{0.6}]^2)$.

I.2.2 MAF

We have used the *tensorflow_probability* library to implement MAF. More specifically, we stacked 8 *shift_and_scale* MAF blocks with $[512, 512]$ hidden layers, each followed by an order flipping permutation layer. We used the same annealing designs discussed above.

I.2.3 Datasets

Table 2 summarizes the datasets used in our Bayesian Logistic regression experiment.

I.3 Reinforcement learning

In the RL experiment, we consider the *swimmer* problem from *rllab*. We used a three layer feed-forward net as our policy net, both hidden layers has 128 units. For the Q-net, we also used a three layer 128 hidden unit feedforward net. we applied a four block *shift_and_scale* MAF as our proposal density, each has two hidden layers with size 512. We set batch size to 128, learning rate to 0.0004 and maintained a 10^6 replay buffer.

References

Ambrosio, L., Gigli, N., and Savaré, G. (2008). *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media.

Jordan, R., Kinderlehrer, D., and Otto, F. (1998). The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17.

Liu, Q. (2017). Stein variational gradient descent as gradient flow. In *NIPS*, pages 3118–3126.

Liu, Q. and Wang, D. (2016). Stein variational gradient descent: A general purpose bayesian inference algorithm. In *NIPS*.

Villani, C. (2008). *Optimal transport: old and new*, volume 338. Springer Science & Business Media.