

846 A Appendix / supplemental material

847 A.1 Proofs

848 We start by the proof of Lemma [1](#), concerned with 0-homogeneity of our target $\theta \mapsto h(\theta)$.

849 *Proof of Lemma [1](#)* We prove in fact that for all $c > 0$, $\pi^\theta = \pi^{c\theta}$. This is a consequence of the fact
 850 that if $\theta \mapsto \phi(x, \theta)$ is 1-homogeneous, then so does $\theta \mapsto C_{\mu\nu}^\phi(\theta)$. Thus, the level set of $\pi \mapsto g(\pi, c\theta)$
 851 are the level-set of $\pi \mapsto g(\pi, \theta)$ dilated by a factor c . Hence, they share the same minimizers, and in
 852 consequence $h(c\theta) = h(\theta)$. \square

853 For the sake of completeness, we also prove the comment stating that a gradient flow on h preserves
 854 the norm of the initialization with the following lemma

855 **Lemma 3.** *Let $U \subseteq \mathbb{R}^q$ an open set, assume that $h : U \mapsto \mathbb{R}$ is differentiable and h is 0-homogeneous,*
 856 *i.e., $h(c\theta) = h(\theta)$. Consider the gradient flow dynamics*

$$\begin{cases} \theta(0) &= \theta_0 \in U \\ \dot{\theta}(t) &= -\nabla h(\theta(t)). \end{cases} \quad (9)$$

857 *Then, there exists an interval $I \subseteq \mathbb{R}_{\geq 0}$ and a unique solution $t \mapsto \theta(t)$ such that for all $t \in I$,*
 858 *$\langle \dot{\theta}(t), \theta(t) \rangle = 0$ and $\|\theta(t)\| = \theta_0$.*

859 *Proof. Orthogonal gradient of 0-homogeneous function.* Consider the real function $\psi : \mathbb{R} \rightarrow \mathbb{R}$
 860 defined by $\psi(c) = h(c\theta)$. By 0-homogeneity, $\psi(c) = h(\theta) = \psi(1)$. Hence, ψ is constant on \mathbb{R}^* ,
 861 thus $\psi'(c) = 0$ for all $c \neq 0$. Using the chain rule for real function, we have that for all $c \neq 0$

$$\psi'(c) = \langle (c \mapsto c\theta)'(c), \nabla h(c\theta) \rangle = \langle \theta, \nabla h(c\theta) \rangle = 0.$$

862 Using this fact for $c = 1$, we conclude that

$$\forall \theta \in U, \quad \langle \theta, \nabla h(\theta) \rangle = 0. \quad (10)$$

863 *Orthogonal dynamics.* The existence and uniqueness of the Cauchy problem [\(9\)](#) comes from the
 864 Cauchy-Lipschitz theorem. Consider this solution $t \mapsto \theta(t)$ defined over I . Then,

$$\langle \dot{\theta}(t), \theta(t) \rangle = -\langle \nabla h(\theta(t)), \theta(t) \rangle = 0,$$

865 using [\(10\)](#). Hence, $\dot{\theta}(t) \perp \theta(t)$ for all $t \in I$.

866 *Conservation of the norm.* Consider $r(t) = \|\theta(t)\|^2$. The chain rule tells us that for all $t \in I$,

$$r'(t) = 2\langle \dot{\theta}(t), \theta(t) \rangle,$$

867 hence $r' = 0$ and thus r is a constant function. \square

868 The proof of Proposition [1](#) relies on the fact that the p -Wasserstein distance is a metric on 1D
 869 measures, and that ϕ^θ is conveniently supposed to be injective over the reference set \mathcal{X} .

870 *Proof of Proposition [1](#)* Let $p > 1$, $\theta \in \mathbb{R}^q$, and assume that ϕ^θ is an injective map from $\mathbb{X} \mathcal{X}?$ to \mathbb{R} .
 871 Let μ, ν, ξ three discrete distributions in $\mathcal{P}(\mathcal{X})$. Recall that

$$d^\theta(\mu, \nu) = W_p(\phi_\#^\theta \mu, \phi_\#^\theta \nu).$$

872 Using the fact that W_p is a metric on $\mathcal{P}_p(\mathbb{R})$, we also obtain that W_p is a metric on $\mathcal{P}(\mathcal{X})$ as a
 873 restriction.

874 **Well-posedness.** Since $d^\theta(\mu, \mu) = W_p(\phi_\#^\theta \mu, \phi_\#^\theta \mu)$, and that W_p is a metric, we have that $d^\theta(\mu, \mu) =$
 875 0 .

876 **Symmetry.** The symmetry comes directly from the one of W_p .

877 **Positivity.** Suppose that $\mu \neq \nu$. Since W_p is a metric, we only need to prove that $\phi_\#^\theta \mu \neq \phi_\#^\theta \nu$.
 878 Assuming that, where $x_i \neq x_{i'}$ for all $i \neq i'$ and $y_j \neq y_{j'}$ for all $j \neq j'$,

$$\mu = \sum_{i=1}^n a_i \delta_{x_i} \quad \text{and} \quad \nu = \sum_{j=1}^m b_j \delta_{y_j},$$

879 we have that

$$\phi_\#^\theta \mu = \sum_{i=1}^n a_i \delta_{\phi^\theta(x_i)} \quad \text{and} \quad \phi_\#^\theta \nu = \sum_{j=1}^m b_j \delta_{\phi^\theta(y_j)}.$$

880 Using the injectivity of ϕ^θ , we have that $\phi^\theta(x_i) \neq \phi^\theta(x_{i'})$ for all $i \neq i'$ and $\phi^\theta(y_j) \neq \phi^\theta(y_{j'})$ for all
 881 $j \neq j'$. Hence, $\phi_\#^\theta \mu = \phi_\#^\theta \nu$ if, and only if, $n = m$ and there exists a permutation $\sigma : \{1, \dots, n\} \rightarrow$
 882 $\{1, \dots, n\}$ such that

$$a_i = b_{\sigma(i)} \quad \text{and} \quad \phi^\theta(x_i) = \phi^\theta(y_{\sigma(i)}), \quad \text{for all } i.$$

883 But then, using the injectivity of ϕ^θ , we have $\{x_i\}_{i=1}^n = \{y_i\}_{i=1}^n$. Hence, $\mu = \nu$ which is a
 884 contradiction.

885 **Triangle inequality.** We have that $d^\theta(\mu, \nu) = W_p(\phi_\#^\theta \mu, \phi_\#^\theta \nu)$. Using the triangle inequality on W_p ,
 886 we have that $d^\theta(\mu, \nu) \leq W_p(\phi_\#^\theta \mu, \phi_\#^\theta \xi) + W_p(\phi_\#^\theta \xi, \phi_\#^\theta \nu) = d^\theta(\mu, \xi) + d^\theta(\xi, \nu)$. \square

887 We now turn to the Stein's lemma. The proof of the Stein's lemma under a (weak) differentiability cri-
 888 terion [51] is classic, and relies on an integration by part and the properties of the normal distribution.
 889 Nevertheless, we are concerned with a function $\theta \mapsto h(\theta)$ that typically will have discontinuities,
 890 breaking the classical proof. Note that one cannot expect the Stein's lemma to hold true for any
 891 kind of discontinuities, even with almost everywhere differentiability. The celebrated example is the
 892 Heaviside function $h(\theta) = 1_{\theta \geq 0}$ in 1D where the Stein's lemma needs a correction term if there is a
 893 non-negligible number of them in sense of the \mathcal{H}^{q-1} Hausdorff dimension. This setting was studied
 894 by [25, 53, 24] for various applications in statistics and signal processing, in particular for Stein's
 895 unbiased risk estimation.

896 *Proof of Lemma 2* The proof of this result is mostly contained in [24, Proposition 1]. We outline
 897 the strategy. Assumption 1 requires to have $\mathcal{H}^{q-1}(\mathbb{R}^q \setminus C) = 0$. Hence, for all $i \in \{1, \dots, q\}$ and
 898 Lebesgue almost all $(\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_q) \in \mathbb{R}^{q-1}$, the map

$$t \mapsto h(\theta_1, \dots, \theta_{i-1}, t, \theta_{i+1}, \dots, \theta_q)$$

899 is absolutely continuous on every compact interval of \mathbb{R} . So, in turns, $h \in W_{\text{loc}}^{1,1}(\mathbb{R}^q)$ which in turns
 900 show that h is almost differentiable in the sense of Stein [51] and we can thus apply [51, Lemma
 901 1]. \square

902 We now turn to the proof of Proposition 2 regarding the properties of the smoothed version h_ε of h .

903 *Proof of Proposition 2 (Admissibility.)* Let $\varepsilon > 0$ and $\theta \in \mathbb{R}^q$. Recall that

$$\pi_\varepsilon^\theta = \mathbb{E}_{Z \sim \mathcal{N}(0, I_q)} \left[\arg \min_{\pi \in U(a, b)} g(\pi, \theta + \varepsilon Z) \right].$$

904 Denote by $u(z) = \arg \min_{\pi \in U(a, b)} g(\pi, \theta + \varepsilon z)$, hence $\pi_\varepsilon^\theta = \mathbb{E}_{Z \sim \mathcal{N}(0, I_q)} [u(Z)]$. For all $z \in \mathbb{R}^q$,
 905 $u(z) \in U(a, b)$ by definition of the minimization problem. Hence,

$$\pi_\varepsilon^\theta = \int_{\mathbb{R}^q} u(z) \rho(z) dz,$$

906 where $\rho(z) = (2\pi)^{-q/2} \exp(-\frac{1}{2}\|z\|^2)$ is the probability density function of the multivariate normal
 907 law and dz is the Lebesgue measure on \mathbb{R}^q . Since $U(a, b)$ is convex and $u(z) \in U(a, b)$, then
 908 $\int_{\mathbb{R}^q} u(z) \rho(z) dz \in U(a, b)$ also. In turn, since $\pi_\varepsilon^\theta \in U(a, b)$, then given the true solution π_{OT}^* of (1),
 909 we have $\sum_{i=1}^n \sum_{j=1}^m \pi_{\text{OT}, i, j}^* \|x_i - y_j\|_p^p \leq \sum_{i=1}^n \sum_{j=1}^m \pi_{i, j}^\theta \|x_i - y_j\|_p^p$

910 (*Differentiability.*) This is a direct consequence of Lemma 2 and Lebesgue dominated convergence
 911 theorem to invert expectation and derivative.

(Consistency.) The first fact is a consequence of Lebesgue dominated convergence theorem. The second one use the expression of the gradient through the variance reduction expression:

$$\nabla_{\theta} h_{\varepsilon}(\theta) = \varepsilon^{-1} \mathbb{E}_Z [(h(\theta + \varepsilon Z) - h(\theta))Z] = \mathbb{E}_Z \left[\frac{h(\theta + \varepsilon Z) - h(\theta)}{\varepsilon} Z \right].$$

Hence, again using Lebesgue dominated convergence theorem, we have

$$\lim_{\varepsilon \rightarrow 0} \nabla_{\theta} h_{\varepsilon}(\theta) = \lim_{\varepsilon \rightarrow 0} \mathbb{E}_Z \left[\frac{h(\theta + \varepsilon Z) - h(\theta)}{\varepsilon} Z \right] = \mathbb{E}_Z \left[\lim_{\varepsilon \rightarrow 0} \frac{h(\theta + \varepsilon Z) - h(\theta)}{\varepsilon} Z \right].$$

Recognizing the directional derivative of $h(\theta)$ if h is differentiable at θ , we get that

$$\lim_{\varepsilon \rightarrow 0} \nabla_{\theta} h_{\varepsilon}(\theta) = \mathbb{E}_Z [\langle \nabla h(\theta), Z \rangle Z] = \nabla h(\theta).$$

(Distance.) We assume here that ϕ^{θ} is injective on \mathcal{X} . We split the proof for h (1.) and h_{ε} (2.).

1. Proof that $(\mu, \nu) \mapsto h(\theta)(\mu, \nu)$ is a distance over $\mathcal{P}(\mathcal{X})$. The *positivity* comes from the suboptimality of $\pi^{\theta}(\mu, \nu)$, that is $h(\theta)(\mu, \nu) \geq W_p^p(\mu, \nu) > 0$ if $\mu \neq \nu$ (as W_p is a metric itself). The *symmetry* comes from the fact that $C_{\mu\nu}$ is symmetric and that $\pi^{\theta}(\mu, \nu) = (\pi^{\theta}(\nu, \mu))^{\top}$. Regarding the *well-posedness*, since ϕ^{θ} is injective, then $W_p(\phi_{\#}^{\theta}\mu, \phi_{\#}^{\theta}\mu) = 0$ and $\pi^{\theta}(\mu, \mu)$ is the identity matrix. Hence,

$$h(\theta)(\mu, \mu) = \langle C_{\mu\mu}, \pi^{\theta}(\mu, \mu) \rangle = \sum_{i=1}^n \sum_{j=1}^n (C_{\mu\mu})_{ij} \pi_{ij}^{\theta}(\mu, \mu) = \sum_{i=1}^n (C_{\mu\mu})_{ii} = 0,$$

since for all i , $(C_{\mu\mu})_{ii} = \|x_i - x_i\|_p^p = 0$. Concerning the *triangle inequality*, let $\mu_1, \mu_2, \mu_3 \in \mathcal{P}(\mathcal{X})$. Let us denote

$$\pi^{12} = \pi^{\theta}(\mu_1, \mu_2) \in \mathbb{R}^{n_1 \times n_2}, \quad \pi^{13} = \pi^{\theta}(\mu_1, \mu_3) \in \mathbb{R}^{n_1 \times n_3}, \quad \pi^{23} = \pi^{\theta}(\mu_2, \mu_3) \in \mathbb{R}^{n_2 \times n_3}$$

Using the specific structure of the 1D optimal transport [50], there exists a tensor $\Pi \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ of order 3 such that admits π^{12} , π^{13} and π^{23} as marginals, that is

$$\begin{aligned} \forall i, j, \quad \pi_{i,j}^{12} &= \sum_{k=1}^{n_3} \Pi_{i,j,k} \\ \forall i, k, \quad \pi_{i,k}^{13} &= \sum_{j=1}^{n_2} \Pi_{i,j,k} \\ \forall j, k, \quad \pi_{j,k}^{23} &= \sum_{i=1}^{n_1} \Pi_{i,j,k}. \end{aligned}$$

Since this structure provides us a “gluing lemma”, we continue the proof similarly to the standard proof of the triangular inequality of the Wasserstein distance.

$$\begin{aligned} (h(\theta)(\mu_1, \mu_3))^{1/p} &= (\langle C_{\mu_1\mu_3}, \pi^{13} \rangle)^{1/p} \\ &= \left(\sum_{i=1}^{n_1} \sum_{k=1}^{n_3} \pi_{i,k}^{13} \|x_i - z_k\|_p^p \right)^{1/p} && \text{by definition} \\ &= \left(\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \Pi_{i,j,k} \|x_i - z_k\|_p^p \right)^{1/p} && \text{as glue.} \end{aligned}$$

Using that $\|x_i - z_k\|_p^p \leq \|x_i - y_j\|_p^p + \|y_j - z_k\|_p^p$, we get that

$$(h(\theta)(\mu_1, \mu_3))^{1/p} \leq \left(\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \Pi_{i,j,k} (\|x_i - y_j\|_p^p + \|y_j - z_k\|_p^p) \right)^{1/p}.$$

Applying now the Minkowski inequality, we obtain that

$$(h(\theta)(\mu_1, \mu_3))^{1/p} \leq \left(\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \Pi_{i,j,k} \|x_i - y_j\|_p^p \right)^{1/p} + \left(\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \Pi_{i,j,k} \|y_j - z_k\|_p^p \right)^{1/p}.$$

930 Using the fact that Π has marginals π^{12} and π^{23} , we get that

$$(h(\theta)(\mu_1, \mu_3))^{1/p} \leq \left(\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \pi_{ij}^{12} \|x_i - y_j\|_p^p \right)^{1/p} + \left(\sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \pi_{jk}^{23} \|y_j - z_k\|_p^p \right)^{1/p}.$$

931 Hence,

$$(h(\theta)(\mu_1, \mu_3))^{1/p} \leq (h(\theta)(\mu_1, \mu_2))^{1/p} + (h(\theta)(\mu_2, \mu_3))^{1/p}.$$

932 **2.** Proof that $(\mu, \nu) \mapsto h_\varepsilon(\theta)(\mu, \nu)$ is a distance over $\mathcal{P}(\mathcal{X})$. The *well-posedness* comes from the fact
 933 that $h_\varepsilon(\theta)(\mu, \mu) = \mathbb{E}_Z[h(\theta + \varepsilon Z)(\mu, \mu)] = \mathbb{E}_Z[0] = 0$. The symmetry is also a direct consequence
 934 of the *symmetry* of $h(\theta)(\mu, \nu)$ and the *positivity* comes from the fact that the expectation of a positive
 935 quantity is positive (understood almost surely on \mathbb{R}^q). For the *triangle inequality*, we use the linearity
 936 of the expectation: let $\mu_1, \mu_2, \mu_3 \in \mathcal{P}(\mathcal{X})$. Then, for all $\theta \in \mathbb{R}^q$, and for all $z \in \mathbb{R}^q$, using the fact
 937 that $h(\theta + \varepsilon z)$ is a distance

$$h(\theta + \varepsilon z)(\mu_1, \mu_3)^{1/p} \leq h(\theta + \varepsilon z)(\mu_1, \mu_2)^{1/p} + h(\theta + \varepsilon z)(\mu_2, \mu_3)^{1/p}.$$

938 Hence, taking the expectation and using linearity gives that

$$\mathbb{E}_Z[h(\theta + \varepsilon Z)(\mu_1, \mu_3)^{1/p}] \leq \mathbb{E}_Z[h(\theta + \varepsilon Z)(\mu_1, \mu_2)^{1/p}] + \mathbb{E}_Z[h(\theta + \varepsilon Z)(\mu_2, \mu_3)^{1/p}].$$

939

□

940 A.2 Algorithm

941 Algorithm 1 describes a gradient descent method to perform the minimization of h_ε using the
 942 Monte-Carlo approximation from Eq. (8).

Algorithm 1 Monte-Carlo gradient descent of $h_\varepsilon(\theta)$

Require: $\theta_0 \in \mathbb{R}^q$, step size policy η_t , smoothing parameter $\varepsilon > 0$, number of Monte Carlo samples N , number of iterations T

- 1: **for** $t = 0$ to $T - 1$ **do**
 - 2: Sample i.i.d. perturbation vectors $z_1, \dots, z_N \sim \mathcal{N}(0, \text{Id}_q)$
 - 3: **for** $k = 1$ to N **do**
 - 4: Solve OT problems to obtain $\pi^{\theta_t + \varepsilon z_k}$ and π^{θ_t} ▷ using 1D OT solver
 - 5: $g_k \leftarrow \langle C_{\mu\nu}, \pi^{\theta_t + \varepsilon z_k} - \pi^{\theta_t} \rangle$
 - 6: $\hat{\nabla} h_{\varepsilon, N}(\theta_t) \leftarrow \frac{1}{\varepsilon N} \sum_{k=1}^N g_k z_k$ ▷ approximate gradient
 - 7: $\theta_{t+1} \leftarrow \theta_t - \eta_t \hat{\nabla} h_{\varepsilon, N}(\theta_t)$ ▷ update parameter
 - 8: **return** θ_T
-

943 A.3 Additional results and experiment details

944 All experiments except the Conditional Flow Matching (CFM) were run on a MacBook Pro M2 Max
 945 with 32 GB of RAM. On this machine, Fig. 1 took approximately 3 minutes per run (10,000 iterations),
 946 Fig. 3 about 6 minutes for 10 runs (with two models trained sequentially, 1,000 iterations), Fig. 4
 947 required roughly 30 minutes, and Fig. 5 took around 10 minutes in total (all models considered, 10
 948 repetitions). The CFM experiments were dispatched over a GPU cluster composed of GPU-A100 80G,
 949 GPU-A6000 48 Go, with a total runtime of 130h for training and inference of all presented models.
 950 We estimate that the total compute time over the course of the project—including experimentation,
 951 debugging, and hyperparameter tuning—is approximately two orders of magnitude larger than the
 952 reported runtimes for CPU-based experiments, and one order of magnitude larger for the GPU-based
 953 experiments.

954 A.3.1 Hyperparameter settings

955 We report here the hyperparameter configurations used across the main experiments. Figures 1 and 3
 956 correspond to the same experiment—Fig. 3 highlights early training dynamics, while Fig. 1 depicts
 957 results at convergence. The projection network used is a 3-layer MLP with ReLU activations: (with



Figure 7: Impact of the variance reduction scheme from Eq. 8. Here, the same learning rate is used for both variants, in which case the variant without variance reduction does not even converge.

dimensions $2 \rightarrow 64 \rightarrow 16 \rightarrow 1$). Optimization is done using SGD with a learning rate of 0.2; for the variant without variance reduction, a lower learning rate of 0.0002 is used to ensure convergence (cf. Fig. 7 in which the same learning rate is used for both variants). In Figure 4 (gradient flow experiments), we perform 2000 outer flow steps using SGD with a learning rate of 0.01. At each flow step, we execute 20 projection steps (or inner optimization updates when using learnable projectors). For the latter, we use Adam with a learning rate of 0.01. The neural projector for our method is a single-hidden-layer MLP with ReLU activations and He initialization. In Figure 5 which investigates gradient flows on hyperbolic manifolds, we vary the outer learning rate across methods to account for differences in convergence speed: the base learning rate is 2.5, used for HHSW; SW uses a scaled learning rate of 17.5, and DGSWP uses a reduced rate of 0.83. Each flow step is composed of 100 projection or inner optimization steps. For the Conditional Flow Matching (CFM) experiment shown in Figures 6, 8, 9 and Table 1, we adopt the same training hyperparameters as in Tong et al. [54]. For our method specifically, the projection model is a 3-layer fully connected network with SELU activations: $3 \times 32 \times 32 \rightarrow 256 \rightarrow 256 \rightarrow 1$. Its parameters are optimized using Adam with a learning rate of 0.01. We perform 1000 optimization steps for the projection model at initialization, followed by 1 step per CFM training iteration.

A.3.2 Additional results

Fig. 8 presents a set of generated images using I-CFM, OT-CFM and DGSWP (NN) after 400k iterations using the 100-step Euler integrator.

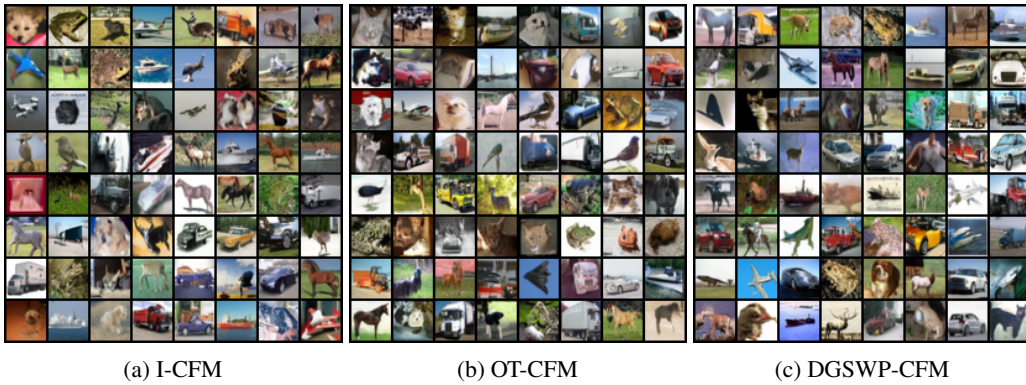


Figure 8: Example of generated images after 400k steps.

Fig. 9 presents the evolution of the image generation quality during training when using the adaptive-step Dormand-Prince strategy for the integration.

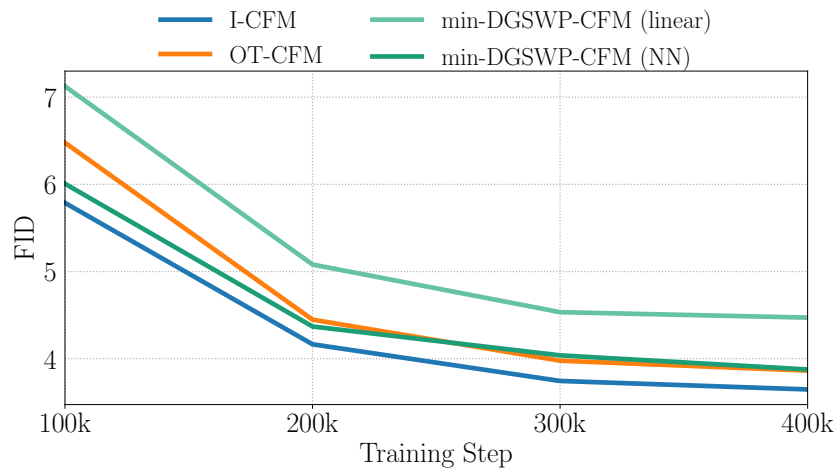


Figure 9: FID as a function of training iterations for various algorithms using adaptive-step DoPri5 sampling.