A The Other Usage

We introduce the usage that inserts a fixed number of gradient layers into the bottom of the generator to assist overall training procedure, which is described in Algorithm 1. Note that we always use latest parameters of f, g for gradient layers in Algorithm 1. When gradient layers are inserted in the middle of the generator: $g_1 \circ \phi \circ g_2$, we can apply Algorithm 1 by setting $\mu_n \leftarrow g_{2\sharp}\mu_n, g \leftarrow g_1$. After training, we can generate samples by using parameters of the critic and the generator, the learning rate, and the number of gradient layers, which is described in Algorithm2.

Algorithm 1 Assisting WGAN-GP

Input: The base distribution μ_n, the minibatch size b, the number of iterations T, the initial parameters τ_0 and θ_0 of the critic and the generator, the number of iterations T_0 for the critic, the regularization parameter c, learning rate η for gradient layers, the number of gradient layers l.

for $k = 0$ **to** $T - 1$ **do**
  $\tau \leftarrow \tau_k$
  for $k_0 = 0$ **to** $T_0 - 1$ **do**
    $\{x_i\}_{i=1}^b \sim \mu_D^b$, $\{z_i\}_{i=1}^b \sim \mu_n^b$, $\{\epsilon_i\}_{i=1}^b \sim U[0,1]^b$
    # $G_\eta^{\tau_k,\theta_k}$ is applied l times.
    $\{z_i\}_{i=1}^b \leftarrow \{g_{\theta_k} \circ G_\eta^{\tau_k,\theta_k} \circ \cdots \circ G_\eta^{\tau_k,\theta_k}(z_i)\}_{i=1}^b$
    $\{\tilde{x}_i\}_{i=1}^b \leftarrow \{\epsilon_i x_i + (1-\epsilon_i) z_i\}_{i=1}^b$
    $v = \nabla_\tau \frac{1}{b} \sum_{i=1}^b [f_\tau(z_i) - f_\tau(x_i) + \lambda R_{f_\tau}(\tilde{x}_i)])$
    $\tau \leftarrow \mathcal{A}(\tau, v)$
  end for
  $\tau_{k+1} \leftarrow \tau$
  $\{z_i\}_{i=1}^b \sim \mu_n^b$
  # $G_\eta^{\tau_{k+1},\theta_k}$ is applied l times.
  $\{z_i\}_{i=1}^b \leftarrow \{G_\eta^{\tau_{k+1},\theta_k} \circ \cdots \circ G_\eta^{\tau_{k+1},\theta_k}\}_{i=1}^b$
  $v \leftarrow -\nabla_\theta \frac{1}{b} \sum_{i=1}^b f_{\tau_{k+1}}(g_{\theta_k}(z_i))$
  $\theta_{k+1} \leftarrow \mathcal{A}(\theta_k, v)$
end for
Return τ_T, θ_T.

Algorithm 2 Data Generation for Algorithm 1

Input: the seed drawn from the base measure $z \sim \mu_n$, the parameter τ and θ of the critic and the generator, the learning rate η, the number of gradient layers l.

Apply gradient layers l times $z' \leftarrow G_\eta^{\tau,\theta} \circ \cdots \circ G_\eta^{\tau,\theta}(z)$
Return the sample $g_\theta(z')$.

We next briefly review Algorithm 1 in which a fixed number of gradient layers with latest parameters is inserted in the bottom of a generator of WGAN-GP. That is, gradient layers modify a noise distribution μ_n to improve the quality of a generator by the functional gradient method.

B Brief Review of Wasserstein Distance

We introduce some facts concerning the Wasserstein distance, which is used for the proof of Proposition 2. We first describe a primal form of the Wasserstein distance. For $p \geq 1$ let $\mathcal{P}_p$ be the set of Borel probability measures with finite p-the moment on $\mathcal{X} \subset \mathbb{R}^v$. For $\mu, \nu \in \mathcal{P}_p$ a probability measure γ on $\mathcal{X} \times \mathcal{X}$ satisfying $\pi^1_\sharp \gamma = \mu$ and $\pi^2_\sharp \gamma = \nu$ is called a *plan* (*coupling*), where π^i denotes the projection from $\mathcal{X} \times \mathcal{X}$ to the i-th space $\mathcal{X}$. We denote by $\Gamma(\mu, \nu)$ the set of all plans between μ and ν. We now introduce Kantorovich's formulation

of the p-Wasserstein distance W_p for $p \geq 1$.

$$W_p^p(\mu, \nu) = \min_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} \|x - y\|_2^p d\gamma(x, y) \tag{1}$$

When $p = 1$ and μ, ν have bounded supports, there is the Kantorovich-Rubinstein dual formulation of the 1-Wasserstein distance, which coincide with the definition introduced in the paper. The existence of optimal plans is guaranteed under more general integrand (c.f. [6, 2]) and we denote by Γ the set of optimal plans. Prior to this formulation, the optimal transport problem in Monge's formulation was proposed.

$$\inf_{\phi_\sharp \mu = \nu} \int_{\mathcal{X}} \|x - \phi(x)\|_2^p d\mu(x), \tag{2}$$

where the infimum is taken over all transport maps $\phi : \mathcal{X} \to \mathcal{X}$ from μ to ν, i.e., $\phi_\sharp \mu = \nu$. Because a transport map ϕ gives a plan $\gamma = (id \times \phi)_\sharp \mu$, we can easily find (1) $\leq$ (2). In general, an optimal transport map that solves the problem (2) does not always exist unlike Kantrovich problem (1). However, in the case where $p > 1$, $\mathcal{X} = \mathbb{R}^v$, and μ is absolutely continuous with respect to the Lebesgue measure, the existence of optimal transport maps is guaranteed [3, 4] and it is extended to more general integrand (see [2]). Moreover, this optimal transport map also solves Kantrovich problem (1), i.e., these two distances coincide. On the other hand, in the case $p = 1$, the existence of optimal transport maps is much more difficult, but it is shown in limited settings as follows.

Proposition A (Sudakov [5], see also [1])**.** *Let $\mathcal{X}$ be a compact convex subset in $\mathbb{R}^v$ and assume that μ is absolutely continuous with respect to Lebesgue measure. Then, there exists an optimal transport map ϕ from μ to ν for the problem 2 with $p = 1$. Moreover, if ν is also absolutely continuous with respect to Lebesgue measure, we can choose ψ so that ψ^{-1} is well defined μ_0-a.e., and $\phi_\sharp^{-1}\nu = \mu$.*

Under the same assumption in Proposition A, it is known that two distances (2) and (1) coincide [1], that is, the Kantrovich problem (1) is solved by an optimal transport map.

C Proofs

We here the give proof of Proposition 1.

Proof of Proposition 1. Note that $\mathcal{L}(\psi) = \hat{\mathcal{L}}(f_\psi^*, \psi)$. For $\psi \in B_r^\infty(\phi)$, we divide $\mathcal{L}(\psi)$ into two terms as follows.

$$\mathcal{L}(\psi) = (\hat{\mathcal{L}}(f_\psi^*, \psi) - \hat{\mathcal{L}}(f_\phi^*, \psi)) + \hat{\mathcal{L}}(f_\phi^*, \psi). \tag{3}$$

We first bound the first term in (3) by L-smoothness of $\hat{\mathcal{L}}(f_{\psi'}^*, \psi)$ with respect to ψ' at ψ in $B_r^\infty(\psi)$.

$$\left|\hat{\mathcal{L}}(f_\phi^*, \psi) - (\hat{\mathcal{L}}(f_\psi^*, \psi) + \langle \nabla_{\psi'} \hat{\mathcal{L}}(f_{\psi'}^*, \psi)\big|_{\psi'=\psi}, \phi - \psi\rangle_{L^2(\mu_g)})\right| \leq \frac{L}{2}\|\phi - \psi\|_{L^2(\mu_g)}^2.$$

Since $\hat{\mathcal{L}}(f_{\psi'}^*, \psi)$ attains the maximum, we have $\nabla_{\psi'} \hat{\mathcal{L}}(f_{\psi'}^*, \psi)\big|_{\psi'=\psi} = 0$ and have

$$\left|\hat{\mathcal{L}}(f_\phi^*, \psi) - \hat{\mathcal{L}}(f_\psi^*, \psi)\right| \leq \frac{L}{2}\|\phi - \psi\|_{L^2(\mu_g)}^2. \tag{4}$$

We next bound $\hat{\mathcal{L}}(f_\phi^*, \psi)$ in (3). We remember that

$$\hat{\mathcal{L}}(f_\phi^*, \psi) = \mathbb{E}_{x\sim\mu_D}[f_\phi^*(x)] - \mathbb{E}_{x\sim\mu_g}[f_\phi^* \circ \psi(x)] - \lambda R_{f_\phi^*}. \tag{5}$$

By L-smoothness of f_ϕ^*, it follows that

$$\left|f_\phi^*(\psi(x)) - (f_\phi^*(\phi(x)) + \langle \nabla_z f_\phi^*(z)\big|_{z=\phi(x)}, \psi(x) - \phi(x)\rangle_2)\right| \leq \frac{L}{2}\|\psi(x) - \phi(x)\|_2^2.$$

By taking the expectation with respect to $\mathbb{E}_{\mu_g}$, we get

$$\left|-\mathbb{E}_{x\sim\mu_g}[f_\phi^*\circ\psi(x)]+\mathbb{E}_{\mu_g}[f_\phi^*(\phi(x))]+\langle\nabla_z f_\phi^*\circ\phi,\psi-\phi\rangle_{L^2(\mu_g)}\right|\leq\frac{L}{2}\|\psi-\phi\|^2_{L^2(\mu_g)}.$$

We substitute this inequality into (5), we have

$$\begin{aligned}\hat{\mathcal{L}}(f_\phi^*,\psi)&\leq\mathbb{E}_{x\sim\mu_D}[f_\phi^*(x)]+\frac{L}{2}\|\psi-\phi\|^2_{L^2(\mu_g)}-(\mathbb{E}_{\mu_g}[f_\phi^*(\phi(x))]+\langle\nabla_z f_\phi^*\circ\phi,\psi-\phi\rangle_{L^2(\mu_g)})-\lambda R_{f_\phi^*}\\&=\hat{\mathcal{L}}(f_\phi^*,\phi)-\langle\nabla_z f_\phi^*\circ\phi,\psi-\phi\rangle_{L^2(\mu_g)}+\frac{L}{2}\|\psi-\phi\|^2_{L^2(\mu_g)}\\&=\mathcal{L}(\phi)+\langle\nabla_\phi\mathcal{L}(\phi),\psi-\phi\rangle_{L^2(\mu_g)}+\frac{L}{2}\|\psi-\phi\|^2_{L^2(\mu_g)},\end{aligned}\tag{6}$$

and the opposite inequality

$$\hat{\mathcal{L}}(f_\phi^*,\psi)\geq\mathcal{L}(\phi)+\langle\nabla_\phi\mathcal{L}(\phi),\psi-\phi\rangle_{L^2(\mu_g)}-\frac{L}{2}\|\psi-\phi\|^2_{L^2(\mu_g)},\tag{7}$$

where we used $\nabla_\phi\mathcal{L}(\phi)=-\nabla_z f_\phi^*(z)|_{z=\phi(\cdot)}$. By combining (3),(4), and (6), we have

$$\mathcal{L}(\psi)\leq\mathcal{L}(\phi)+\langle\nabla_\phi\mathcal{L}(\phi),\psi-\phi\rangle_{L^2(\mu_g)}+L\|\phi-\psi\|^2_{L^2(\mu_g)}.$$

Moreover, since $\hat{\mathcal{L}}(f_\psi^*,\psi)-\hat{\mathcal{L}}(f_\phi^*,\psi)\geq 0$ in (3), we have $\mathcal{L}(\psi)\geq\hat{\mathcal{L}}(f_\phi^*,\psi)$. Therefore, we get the opposite inequality by (7)

$$\mathcal{L}(\psi)\geq\mathcal{L}(\phi)+\langle\nabla_\phi\mathcal{L}(\phi),\psi-\phi\rangle_{L^2(\mu_g)}-\frac{L}{2}\|\phi-\psi\|^2_{L^2(\mu_g)}.$$

This finishes the proof. □

We next provide the proof of Theorem 1.

Proof of Theorem 1. Noting that $\|\eta\nabla_{\phi_k}\mathcal{L}(\phi_k)\|_\infty\leq r$ and Lipschitz smoothness of $\mathcal{L}$, we have

$$\begin{aligned}\mathcal{L}(\phi_{k+1})&\leq\mathcal{L}(\phi_k)-\eta\|\nabla_\phi\mathcal{L}(\phi_k)\|^2_{L^2(\mu_g)}+\frac{\eta^2L}{2}\|\nabla_\phi\mathcal{L}(\phi_k)\|_{L^2(\mu_g)}\\&=\mathcal{L}(\phi_k)-\eta(1-\eta L/2)\|\nabla_\phi\mathcal{L}(\phi_k)\|^2_{L^2(\mu_g)}.\end{aligned}$$

Since $\eta\leq 1/L$, we have $\mathcal{L}(\phi_{k+1})\leq\mathcal{L}(\phi_k)-\frac{\eta}{2}\|\nabla_\phi\mathcal{L}(\phi_k)\|^2_{L^2(\mu_g)}$. Summing up over $k\in\{0,\dots,T-1\}$ and dividing by T we obtain

$$\frac{1}{T}\sum_{k=0}^{T-1}\|\nabla_\phi\mathcal{L}(\phi_k)\|^2_{L^2(\mu_g)}\leq\frac{2}{\eta T}(\mathcal{L}(\phi_0)-\mathcal{L}(\phi_T)).$$

This inequality finishes the proof of the theorem. □

Proof of Proposition 2 . By Proposition A, there exists an optimal transport ψ from μ_g to μ_D and an optimal plan is given by $\gamma=(id\times\psi)_\sharp\mu_g$. We set $\psi_t=(1-t)id+t\psi$ and $\mu_t=\psi_{t\sharp}\mu_g$. Because $(\psi_s,\psi_t)_\sharp\mu_g$ $(0\leq s<t\leq 1)$ gives a plan between μ_g and μ_D, we have

$$\begin{aligned}W_1(\mu_s,\mu_t)&\leq\int_{\mathcal{X}\times\mathcal{X}}\|x-y\|_2 d(\psi_s,\psi_t)_\sharp\mu_g\\&=\int_{\mathcal{X}}\|\psi_s(x)-\psi_t(x)\|_2 d\mu_g\\&=(t-s)\int_{\mathcal{X}}\|x-\psi(x)\|_2 d\mu_g=(t-s)W_1(\mu_g,\mu_D).\end{aligned}\tag{8}$$

We next prove the opposite inequality. Noting that $(id, \psi_s)_\sharp \mu_g$ is a plan from μ_g to μ_s and $(\psi_t, \psi)_\sharp \mu_g$ is a plan from μ_t to μ_D, we have the following two inequalities

$$W_1(\mu_g, \mu_s) \leq \int_{\mathcal{X}\times\mathcal{X}} \|x - y\|_2 d(id, \psi_s)_\sharp \mu_g = \int_{\mathcal{X}} \|x - \psi_s(x)\|_2 d\mu_g = sW_1(\mu_g, \mu_D),$$
$$W_1(\mu_t, \mu_D) \leq \int_{\mathcal{X}\times\mathcal{X}} \|x - y\|_2 d(\psi_t, \psi)_\sharp \mu_g = \int_{\mathcal{X}} \|\psi_t(x) - \psi(x)\|_2 d\mu_g = (1-t)W_1(\mu_g, \mu_D).$$

Using these two inequalities and the triangle inequality, we get

$$W_1(\mu_g, \mu_D) \leq W_1(\mu_g, \mu_s) + W_1(\mu_s, \mu_t) + W_1(\mu_t, \mu_D) \leq (1 + s - t)W_1(\mu_g, \mu_D) + W_1(\mu_s, \mu_t).$$

That is $(t-s)W_1(\mu_g, \mu_D) \leq W_1(\mu_s, \mu_t)$. By combining this inequality and (8), we have $(t-s)W_1(\mu_g, \mu_D) = W_1(\mu_s, \mu_t)$ and this finishes the proof. □

D Labeled Faces in the Wild

In this section we provide the result on the Labeled Faces in the Wild dataset. The result is shown in Figure 1. After training WGAN-GP (left), we ran Algorithm 3 for a few iterations (right).

Figure 1: Random samples drawn from the generator trained by WGAN-GP (left) and the gradient layer (right).

References

[1] Luigi Ambrosio. Lecture notes on optimal transport problems. In *Mathematical aspects of evolving interfaces*, pages 1–52. Springer, 2003.

[2] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[3] Yann Brenier. Décomposition polaire et réarrangement monotone des champs de vecteurs. *CR Acad. Sci. Paris Sér. I Math*, 305(19):805–808, 1987.

[4] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

[5] Vladimir N Sudakov. *Geometric problems in the theory of infinite-dimensional probability distributions*. Number 141. American Mathematical Soc., 1979.

[6] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.