Gradient Layer: Enhancing the Convergence of Adversarial Training for Generative Models

Atsushi Nitanda†1,2
1Graduate School of Information Science and Technology, The University of Tokyo
2Center for Advanced Intelligence Project, RIKEN

Taiji Suzuki‡1,2,3
3PRESTO, Japan Science and Technology Agency

Abstract

We propose a new technique that boosts the convergence of training generative adversarial networks. Generally, the rate of training deep models reduces severely after multiple iterations. A key reason for this phenomenon is that a deep network is expressed using a highly non-convex finite-dimensional model, and thus the parameter gets stuck in a local optimum. Because of this, methods often suffer not only from degeneration of the convergence speed but also from limitations in the representational power of the trained network. To overcome this issue, we propose an additional layer called the gradient layer to seek a descent direction in an infinite-dimensional space. Because the layer is constructed in the infinite-dimensional space, we are not restricted by the specific model structure of finite-dimensional models. As a result, we can get out of the local optima in finite-dimensional models and move towards the global optimal function more directly. In this paper, this phenomenon is explained from the functional gradient method perspective of the gradient layer. Interestingly, the optimization procedure using the gradient layer naturally constructs the deep structure of the network. Moreover, we demonstrate that this procedure can be regarded as a discretization method of the gradient flow that naturally reduces the objective function. Finally, the method is tested using several numerical experiments, which show its fast convergence.

1 Introduction

Generative adversarial networks (GANs) [7] are a promising scheme for learning generative models. GANs are trained by a discriminator and a generator in an adversarial way. Discriminators are trained to classify between real samples and fake samples drawn from generators, whereas generators are trained to mimic real samples. Although training GANs is quite difficult, adversarial learning succeeded in generating very impressive samples [19], and there are many subsequent studies [14, 20, 17, 6, 26]. Wasserstein GANs (WGANs) [3] are a variant to remedy the mode collapse that appears in the standard GANs by using the Wasserstein distance [24], although they also sometimes generate low-quality samples or fail to converge. Moreover, an improved variant of WGANs was also proposed [8] and it succeeded in generating high-quality samples and stabilizing WGANs. Although these attempts have provided better results, there is still scope to improve the performance of GANs further.

One reason for this difficulty stems from the limitation of the representational power of the generator. If the discriminator is optimized for the generator, the behavior is solely determined by the samples produced from that generator. In other words, for a generator with a poor representational power, the discriminator terminates its learning in the early stage and consequently results in having low discriminative power. However, for a finite-dimensional parameterized generator, the ability to generate novel samples to cheat the discriminators is limited. In addition, the highly non-convex structure of the deep neural network for the generator prevents us from finding a direction for improvement. As a result, the trained parameter gets stuck in a local optimum and the training procedure does not proceed any more.

In this study, we propose a new learning procedure to overcome the issues of limited representational power...
Gradient Layer: Enhancing the Convergence of Adversarial Training for Generative Models

and local optimum by introducing a new type of layer called a gradient layer. The gradient layer finds a direction for improvement in an infinite-dimensional space by computing the functional gradient \[15\] instead of the ordinary gradient induced by a finite-dimensional model. Because the functional gradient used for the gradient layer is not limited in the tangent space of a finite-dimensional model, it has much more freedom than the ordinary finite-dimensional one. Thanks to this property, our method can break the limit of the local optimum induced by the strong non-convexity of a finite-dimensional model, which gives much more representational power to the generator. We theoretically justify this phenomenon from the functional gradient method perspective and rigorously present a convergence analysis. Interestingly, one iteration of the method can be recognized as inserting one layer into the generator and the total number of iterations is the number of inserted layers. Therefore, our learning procedure naturally constructs the deep neural network architecture by inserting gradient layers. Although, gradient layers can be inserted into an arbitrary layer, they are typically stacked on top of the generator in the final training phase to improve the generated sample quality.

Moreover, we provide another interesting perspective of the gradient layer, i.e., discretization of the gradient flow in the space of probability measures. In Euclidean space, the steepest descent which is the typical optimization method, can be derived by discretizing the gradient flow that naturally produces a curve to reduce the objective function. Because the goal of GANs is to generate a sequence of probability measures moving to the empirical distribution by training samples, it is natural to consider a gradient flow in the space of probability measures defined by a distance between generated distribution and the empirical distribution and to discretize it in order to construct practical algorithms. We show that the functional gradient method for optimizing the generator in the function space is such a discretization method; in other words, the gradient flow can be tracked by stacking gradient layers successively.

Related Work The recently proposed SteinGAN \[25\] is based on another strategy to track gradient flow. However, because that discretization is mimicked by a fixed-size deep neural network, it may have the same limitation as typical GANs. By contrast, our method directly tracks the gradient flow in the final phase of training GANs to break the limit of the finite-dimensional generator.

2 Brief Review of Wasserstein GANs

In this section, we introduce WGANs and their variants. Although our proposed gradient layer is applicable to various models, we demonstrate how it performs well for the training of generative models; in particular, we treat Wasserstein GANs as a main application in this paper. Let us start from briefly reviewing WGANs.

WGAN is a powerful generative model based on the 1-Wasserstein distance, defined as the \(L^1\) minimum cost of transporting one probability distribution to the other. Let \(X \subset \mathbb{R}^n\) and \(Z \subset \mathbb{R}^h\) be a compact convex data space and a hidden space, respectively. A typical example of \(X\) is the image space \([0, 1]^n\). For a noise distribution \(\mu_n\) on \(Z\), WGAN learns a data generator \(g : Z \to X\) to minimize an approximation to the 1-Wasserstein distance between the data distribution \(\mu_D\) and the push-forward distribution \(g_\# \mu_n\), which is a distribution that the random variable \(g(z)\) follows when \(z \sim \mu_n\) (in other words, the distribution obtained by applying a coordinate transform \(g\) to \(z \sim \mu_n\)). That is, WGAN can be described as the following min max problem by using a Kantrovich-Rubinstein dual form of the 1-Wasserstein distance:

\[
\min_{g \in \mathcal{G}} \max_{f \in \mathcal{F}} \mathcal{L}(f, g) \overset{\text{def}}{=} \mathbb{E}_{x \sim \mu_D}[f(x)] - \mathbb{E}_{z \sim \mu_n}[f \circ g(z)],
\]

where \(\mathcal{G}\) is the set of generators and \(\mathcal{F}\) is an approximate set to the set of 1-Lipschitz continuous functions called critic. In WGANs, \(\mathcal{G}, \mathcal{F}\) are parameterized neural networks \(\{g_\theta\}, \{f_\tau\}\) and the problem is solved by alternate optimization: maximizing and minimizing \(\mathcal{L}(f_\tau, g_\theta)\) with respect to \(\tau\) and \(\theta\), alternately.

In practice, to impose the Lipschitz continuity on critics \(f_\tau\), penalization techniques were explored. For instance, the original WGANs \[3\] use weight clipping \(\|\tau\|_\infty \leq c\), which implies the upper-bound on the norm of \(\nabla_\tau f_\tau\) and makes it Lipschitz continuous. However, it was pointed out in a subsequent study \[5\] that such a restriction seems to be unnatural and sometimes leads to a low-quality generator or a failure to converge. In the same study, an improved variant of WGANs called WGAN-GP was proposed, which succeeded in stabilizing the optimization process and generating high-quality samples. WGAN-GP \[8\] adds the gradient penalty \((\|\nabla_\tau f_\tau(\tilde{x})\|_2 - 1)^2\) to the objective function in the training phase of critics, where \(\tilde{x}\) is a random interpolation between a training example \(x \sim \mu_D\) and a generated sample \(g(z) \sim g_\# \mu_n\), i.e., \(\tilde{x} \leftarrow cx + (1-c)g(z) (c \sim U[0, 1]: \text{uniform distribution})\). DRAGAN \[12\] is a similar method to WGAN-GP, although it is based on a different motivation. DRAGAN also uses the gradient penalty, but the penalty is imposed on a neighborhood of the data manifold by a
optimizing $WGAN$ and its variants are learned by alternately random perturbation of a training example. We can regard this learning procedure as a problem of minimizing $L(g_b) = \max_{\tau} \{ L(f_\tau, g_b) - \lambda R_\tau \}$, where $R_\tau$ is a penalty term. Let $L(f_\tau, g_b) - \lambda R_\tau$ attain its maximum value at $\tau_*$ for $g_b$. Then, the gradient $\nabla_g L(g_b)$ is the same as $-E_{\mu_n}[\nabla_\tau f_\tau^1 \nabla_g g_b(z)]$ by the envelope theorem when both terms are well-defined. The differentiability of $L(g_b)$ with respect to $\theta$ almost everywhere is proved in [3] under a reasonable assumption. Hence, we can apply the gradient method to this problem by approximating this gradient with finite particles generated from $\mu_n$. However, because it is difficult to obtain $f_\tau$, we run the gradient method for several iterations on training a critic instead of exactly computing $f_\tau$ at each $g_b$. We can notice that this learning procedure is quite similar to that of the standard GAN [7].

3 Gradient Layer

In the usual training procedure of WGANs, though more general maps are admissible for the original purpose, generators are parameterized by finite-dimensional space as described in the previous section, and the parameter may get stuck in a local optimum induced by this restriction, or the speed of convergence may reduce. In this work, we propose a gradient layer that accelerates the convergence and breaks the limit of finite-dimensional models. This layer is theoretically derived by the infinite-dimensional optimization method. We first explain the high-level idea of the gradient layer that strictly improves the ability of generator and why our method enhances the convergence of training WGANs.

3.1 High-level idea of gradient layer

Here, we explain gradient layer with intuitive motivation. It is inserted into the generator $g$ in WGANs. We now focus on minimizing $L(f, g)$ with respect to $g$ under a fixed critic $f$, that is, we consider the problem $\min_g L_f(g) \overset{\text{def}}{=} E_{\mu_n}[-f(g(z))]$. Let us split $g$ into two neural networks $g = g_1 \circ g_2$ at arbitrary layer where a new layer is to be inserted. Our purpose is to specify the form of layer $\phi$ that reduces the objective value by perturbations of inputs $g_2(z)$, i.e., $L_f(g_1 \circ \phi \circ g_2) \leq L_f(g)$. Since $L_f(g_1 \circ \phi \circ g_2)$ is regarded as the integral $E_{z' \sim g_2 \mu_n}[-f(g_1(\phi'(z')))]$ with respect to the push-forward distribution $g_2 \mu_n$, this purpose is achieved by transporting the input distribution of $\phi$ along the gradient field $\nabla_{z'} f(g_1(z'))$. Therefore, we propose a gradient layer $G_\eta$ with one hyperparameter $\eta > 0$ as a map that transforms an input $z'$ to

$$G_\eta(z') = z' + \eta \nabla_{z'} f(g_1(z')).$$

(1)

Because the gradient layer depends on the parameters $\tau, \theta$ of the upper layers $f, g_1$, we specify the parameter as $G_\eta^{\tau, \theta}$ if needed.

Applying the gradient layer recursively, it further progresses and achieves a better objective. The computation of the gradient layer is quite simple. Actually, simply taking the derivative is sufficient, which can be efficiently executed. Because too many gradient layers would lead to overfitting to the critic $f$, we stop stacking the gradient layer after an appropriate number of steps. Indeed, if $f \circ g_1$ is Lipschitz continuous, $id + \eta f \circ g_1$ for sufficiently small $\eta$ is an injection because $(id + \eta f \circ g_1)(z) = (id + \eta f \circ g_1)(z')$ implies $\|z - z'\| \leq \eta \mathcal{L}_{g_1} \|z - z'\|$ where $\mathcal{L}_{g_1}$ is the Lipschitz constant. Thus, a topology of $\supp(g_2 \mu_n)$ is preserved and early stopping is justified. Then, this layer efficiently generates high-quality samples for the critic and the overall adversarial training procedure can be also boosted.

3.2 Powerful optimization ability

Because the gradient layer directly transports inputs as stated above, it strictly improves the objective value if there is room for optimization, unlike finite-dimensional models that may be trapped in local optima induced by the restriction of generators. Indeed, when the gradient layer cannot move inputs, i.e., $G_\eta(z') = z'$, the gradient $\nabla_{z'} f(g_1(z'))$ vanishes on $\supp(g_2 \mu_n)$ and there is no chance to improve the
objective value by optimizing $g_2$ because of the chain rule of derivatives. We now explain this phenomenon more precisely. Let us first consider the training of $g_2$ in the usual way. We denote by $w_2$ the parameter of $g_2$. As stated in the previous section, $w_2$ is updated by using the gradient

$$E_{\mu_n} [J_{w_2}^T g_2(z) \nabla_z f(g_1(g_2(z)))],$$

where $z'$ is the input to $g_1$ and $J_{w_2} g_2(z)$ is the Jacobian matrix of $g_2$ with respect to $w_2$. We immediately notice that this gradient is the inner product of $J_{w_2} g_2(\cdot)$ and $\nabla_z f(g_1(g_2(\cdot)))$ in $L^2(\mu_n)$-space, and the latter term is the output of the gradient layer. Therefore, when the gradient layer cannot move the inputs, the gradient with respect to $w_2$ also vanishes. However, even if the gradient vanishes, the gradient layer $G_2$ can move the inputs in general. Thus, whereas the optimization of $g_2$ may get stuck in a local optimum or be slowed down in this case, the gradient layer strictly improves the quality of the generated samples for the upper layers, because $\nabla_z f(g_1(g_2(z)))$ does not vanish. This is the reason the gradient layer has a greater optimization ability than finite-dimensional models.

### 3.3 Algorithm description

The overall algorithm is described in this subsection. We adopt WGAN-GP as the base model to which gradient layer is applied. Let us denote by $R_{f_\theta}(\tilde{x})$ a gradient penalty term. In a paper on the improved WGANs [8], the use of a two-sided penalty ($||\nabla_{\tilde{x}} f_\theta(\tilde{x})||_2 - 1)^2$ is recommended. However, we also allow the use of the one-sided variant ($\max(||\nabla_{\tilde{x}} f_\theta(\tilde{x})||_2 - 1, 0))^2$. As for the place in which the gradient layer is inserted, we can propose several possibilities, e.g., inserting the gradient layer into (i) the top and (ii) the bottom of the layers of the generator. The latter usage is described in the appendix.

The first usage is stacking gradient layers on the top of the generator, except for normalization to fine-tune the generator in the final phase. Although a normalization term such as tanh is commonly stacked on generators to bound the output range of the generators, gradient layers are typically applied before the normalization layer. Since tanh is a fixed function, it is no problem to combine tanh with critics by reinterpreting $\mathcal{F}$ and $\mathcal{X}$. The gradient layer directly handles the generated samples, so that it may significantly improve the sample quality. Because the gradient $\nabla_{\tilde{x}} f_\theta(x)$ of the critic with respect to data variables provides the direction to improve the quality of the current generated samples, it is expected that we can obtain better results by tracking the gradient iteratively. To compute the output from the gradient layer for a completely new input, we need to reproduce the computation of the gradient layers, which can be realized by saving the history of the parameters of critics and stacking the gradient layers using these parameters. The concrete procedure is described in Algorithm 1. When executing Algorithm 1, the parameter of $g_\theta$ is fixed, so that the push-forward measure $g_\theta \mu_n$ is treated as a base probability measure and we denote it by $\mu_g$. Because the gradient layers depend on the history of the parameters in this case, we specify the parameter to be used: $G_\tau$. For the parameter $\tau$ and the gradient $v$, we denote by $\mathcal{A}(\tau, v)$ one step of a gradient-based method such as SGD with momentum, Adam [11], and RMSProp [22]. From the optimization perspective, we show that Algorithm 1 can be regarded as an approximation to the functional gradient method. From this perspective, we show fast convergence of the method under appropriate assumptions where the objective function is smooth and the critics are optimized in each loop. This theoretical justification is described later. Although Algorithm 1 has a great optimization ability, applying the algorithm to large models is difficult because it requires the memory to register parameters; thus, we propose its usage for fine-tuning in the final phase of training a WGAN-GP. After the execution of Algorithm 1, we can generate samples by using the history of critics, the learning rate, and the base distribution as described in Algorithm 2.

#### Algorithm 1 Finetuning WGAN-GP

**Input:** The base distribution $\mu_g = g_\theta \mu_n$, the minibatch size $b$, the number of iterations $T$, the initial parameters $\tau_0$ of the critic, the number of iterations $T_0$ for the critic, the regularization parameter $\lambda$, and the learning rate $\eta$ for gradient layers.

```plaintext
for $k = 0$ to $T - 1$
do
  $\tau \leftarrow \tau_k$
  for $k_0 = 0$ to $T_0 - 1$
do
    $\{x_i\}_{i=1}^b \sim U \mathbb{D}_0^b$, $\{z_i\}_{i=1}^b \sim \mu_g$, $\{\epsilon_i\}_{i=1}^b \sim U [0, 1]^b$
    $\{z_i\}_{i=1}^b \leftarrow \{G^\tau_0 \cdots \circ G^\tau_{i-1}(z_i)\}_{i=1}^b$
    $\bar{x}_i \leftarrow \epsilon_i x_i + (1 - \epsilon_i) z_i$
    $u \leftarrow \nabla_{\tau, v} \sum_{i=1}^b [f_\tau(x_i) - f_\tau(z_i) + \lambda R_{f_\theta}(\bar{x}_i)]$
  $\tau \leftarrow \mathcal{A}(\tau, v)$
end for
$\tau_{k+1} \leftarrow \tau$
end for
Return ($\tau_1, \ldots, \tau_T$).
```

#### Algorithm 2 Data Generation for Algorithm 1

**Input:** the seed drawn from base measure $z \sim \mu_g = g_\theta \mu_n$, the history of parameters $\{\tau_k\}_k \tau$, and the learning rate $\eta$ for gradient layers.

Return the sample $G^\tau_\eta \circ \cdots \circ G^\tau_1(z)$.
4 Functional Gradient Method

In this section, we provide mathematically rigorous derivation from the functional gradient method [15] perspective under the Fréchet differentiable (functional differentiable) assumption on $L$. That is, we consider an optimization problem with respect to a generator in an infinite-dimensional space. For simplicity, we focus on the case where the gradient layer is stacked on top of a generator $g$ and we treat $g\circ \mu$, as the base measure $\mu$. Thus, in the following we omit the notation $g$ in $L(f, \phi \circ g)$. Let $L^2(\mu)_{\phi}$ be the space of $L^2(\mu)$-integrable maps from $\mathbb{R}^v$ to $\mathbb{R}^v$, equipped with the $(\cdot, \cdot)_{L^2(\mu)}$-inner product: for $\forall \phi_1, \forall \phi_2 \in L^2(\mu)$,

$$
\langle \phi_1, \phi_2 \rangle_{L^2(\mu)} = \mathbb{E}_{\mu_{\phi}} [\phi_1(z) \top \phi_2(z)].
$$

To learn WGAN-GP, we consider the infinite-dimensional problem:

$$
\min_{\phi \in L^2(\mu)} \max_{f \in F} L(f, \phi) - \lambda R_{f_z},
$$

where $R_{f_z}$ is a gradient penalty term. To achieve this goal, we take a Gâteaux derivative along a given map $v \in L^2(\mu)$, i.e., a directional derivative along $v$. Let us denote $\max_{f \in F} \{L(f, \phi) - \lambda R_f\}$ by $L(\phi)$ and $\arg \max_{f \in F} \{L(f, \phi) - \lambda R_f\}$ by $f^*_\phi$ and the corresponding parameter by $\tau^*_\phi$, i.e., $f^*_\phi = f^*_{\tau^*_\phi}$ for $\phi \in L^2(\mu)$. If every $f \in F$ is Lipschitz continuous and differentiable, we can find that by the envelope theorem and Lebesgue’s convergence theorem this derivative takes the form:

$$
\frac{d}{dt} L(\phi + tv)\big|_{t=0} = -\mathbb{E}_{\mu_{\phi}} [\nabla_x f^*_\phi(x)_{\phi}^\top v(z)].
$$

Therefore, $-\nabla_x f^*_\phi(x)_{x=\phi(x)}$ can be regarded as a Fréchet derivative (functional gradient) in $L^2(\mu)$ and we denote it by $\nabla_{\phi} L(\phi)$, which performs like the usual gradient in Euclidean space. Using this notation, the optimization of $L(\phi)$ can be accomplished by Algorithm 3 which is a gradient descent method in a function space. Because the functional gradient has the form $-\nabla_x f^*_\phi \circ \phi$, each iteration of the functional gradient method with respect to $\phi$ is $\phi \leftarrow \phi - \eta \nabla_{\phi} f^*_\phi \circ \phi = (id + \eta \nabla_x f^*_\phi) \circ \phi$, where $\eta$ is the learning rate. We notice here that this iteration is the composition of a perturbation map $id + \eta \nabla_x f^*_\phi$ and a current map $\phi$ and is nothing but stacking a gradient layer $G^\tau_{\phi}$ on $\phi(z)$. In other words, the functional gradient method with respect to $\phi$, i.e., Algorithm 3 is the procedure of building a deep neural network by inserting gradient layers, where the total number of iterations is the number of layers. Moreover, we notice that if we view $\nabla_x f^*_\phi$ as a perturbation term, this layer resembles that of residual networks [9] which is one of the state-of-the-art architectures in supervised learning tasks.

However, executing Algorithm 3 is difficult in practice because the exact optimization with respect to a critic $f$ to compute $L(\phi)$ is a hard problem. Thus, we need an approximation and we argue that Algorithm 1 is such a method. This point can be understood as follows. Roughly speaking, it maximizes $L(f, \phi)$ with respect to $f$ in the inner loop under fixed $\phi = G^\tau_{\phi} \circ G_{\tau-1} \circ \cdots \circ G^\tau_{\phi}$, to obtain an approximate solution $\tau_{k+1}$ to $\tau$, and minimizes that with respect to $\phi$ in the outer loop by stacking $G^\tau_{\phi}$, which is an approximation to $G^\tau_{\phi}$. Thus, Algorithm 1 is an approximated method, but we expect it to achieve fast convergence owing to the powerful optimization ability of the functional gradient method, as shown later. In particular, it is more effective to apply the algorithm in the final phase of training WGAN-GP to fine-tune it, because the optimization ability of parametric models are limited.

Algorithm 3 Functional Gradient Descent

Input: the initial generator $g$ and the learning rate $\eta$.

$\phi_0 \leftarrow g$

for $k = 0$ to $T - 1$

$\phi_{k+1} \leftarrow \phi_k - \eta \nabla_{\phi} L(\phi_k)$

end for

Return the function $f_T$.

5 Convergence Analysis

Let us provide convergence analysis of Algorithm 3 for the problem of the general form: $\min_{\phi} L(\phi)$. The convergence can be shown in an analogous way to that for the finite-dimensional one. To prove this, we make a smoothness assumption on the loss function. We now describe a definition of the smoothness on a Hilbert space whose counterpart in finite-dimensional space is often assumed for smooth non-convex optimization methods.

Definition 1. Let $h$ be a function on a Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$. We call that $h$ is $L$-smooth at $z$ in $U$ if $h$ is differentiable at $z$ and it follows that $\forall z' \in U$:

$$
|h(z') - h(z) - \langle \nabla_{\mathcal{H}}h(z), z' - z \rangle_{\mathcal{H}}| \leq \frac{L}{2} \|z' - z\|^2_{\mathcal{H}}.
$$

The following definition and proposition provide one condition leading to Lipschitz smoothness of $L$. Let us denote by $\| \cdot \|_{L^\infty(\mu_\phi)}$ the sup-norm $\|\psi\|_{L^\infty(\mu_\phi)} = \sup_{\psi \in \text{supp}(\mu_\phi)} \|\psi(z)\|_2$ and by $B^\infty(\phi)$ a ball of center $\phi$ and radius $r$. Let $L(f, \psi) = L(f, \psi) - \lambda R_f$. In the following we assume $f^*_\phi$ is uniquely defined for $\psi \in L^2(\mu)$ and $L$-smoothness with respect to the input $x$. 


Definition 2. For positive values $r$ and $L$, we call that $\mathcal{L}$ is $(r, L)$-regular at $\phi$ when the following condition is satisfied: For $\forall \psi \in B_{\infty}^{\infty}(\phi)$, $\dot{\mathcal{L}}(f_{\psi}^t, \psi)$ is $L$-smooth at $\psi$ with respect to $\psi'$ in $B_{\infty}^{\infty}(\psi)$.

Proposition 1. If $\mathcal{L}$ is $(r, L)$-regular at $\phi$, then $\mathcal{L}$ is $2L$-smooth at $\phi$ in $B_{\infty}^{\infty}(\psi)$.

We now show the convergence of Algorithm 3. The following theorem gives the rate to converge to the stationary point.

Theorem 1. Let us assume the norm of the gradient $\|\nabla_x f_{\phi}^t(x)\|_2$ is uniformly bounded by $\alpha$ and assume $\mathcal{L}$ is $L$-smooth at $\phi$ in $B_{\infty}^{\infty}(\phi)$ for $\forall \phi \in L^2(\mu_\gamma)$. Suppose we run Algorithm 3 with constant learning rate $\eta \leq \min\{1, L, r/\alpha\}$. Then we have for $T \in \mathbb{Z^+}$

$$
\min_{k\in\{0,...,T-1\}} \|\nabla_{\phi} \mathcal{L}(\phi_k)\|_{L^2(\mu_\gamma)} ^2 \leq \frac{2}{\eta T} (\mathcal{L}(\phi_0) - \mathcal{L}_*)
$$

where $\mathcal{L}_* = \inf_\phi \mathcal{L}(\phi)$.

Note that the convergence rate $O(1/T)$ is the same as the gradient descent method for smooth objective in the finite-dimensional one. This means that even though the optimization is executed in the infinite-dimensional space, we do not suffer from the infinite dimensionality in terms of the convergence.

The following rough argument indicates that Algorithm 3 matches with learning WGANs. Let $W_1$ denote the 1-Wasserstein distance with respect to the Euclidean distance on a compact base space $X \subset \mathbb{R}^d$. The following proposition is immediately shown by combining existing results \cite{1, 21}.

Proposition 2. Let $\mu_\gamma$ be a Borel probability measure on $X$ and assume $\mu_\gamma$ is absolutely continuous with respect to the Lebesgue measure. Then, there exists an optimal transport $\psi$ and it follows that $W_1(\psi_{\#} \mu_\gamma, \mu_D) = (1 - t)W_1(\mu_\gamma, \mu_D)$, where $\psi_t = (1 - t)id + t\psi$.

By this proposition, there exists a curve $\psi_t$ strictly reducing distance, i.e., $dW_1(\psi_{\#} \mu_\gamma, \mu_D)/dt < 0$ if $\mu_\gamma \neq \mu_D$. Because $\mathcal{L}$ approximates $W_1$, it is expected that $d\mathcal{L}(\psi_t)/dt < 0$ when $\mu_\gamma$ differs from $\mu_D$. Noting that $d\mathcal{L}(\psi_t)/dt = \langle \nabla_{\phi} \mathcal{L}(\psi_t), \psi - id \rangle_{L^2(\mu_\gamma)}$, the functional gradient $\nabla_{\phi} \mathcal{L}(\psi_t)$ is not vanishing and the objective $\mathcal{L}$ may be strictly reduced by Algorithm 3.

6 Gradient Flow Perspective

In Euclidean space, the step of the steepest descent method for minimizing problems can be derived by the discretization of the gradient flow $d\gamma(t)/dt = -\nabla_x F(\gamma(t))$ where $F$ is an objective function on Euclidean space. Because our goal is to move $\mu_\gamma$ closer to $\mu_D$, we should consider a gradient flow in the space of probability measures. To make this argument rigorously, we need the continuity equation that characterizes a curve of probability measures and the tangent space where velocities of curves should be contained (c.f., \cite{2}). When these notions are provided, the gradient flow is defined immediately and it is quite natural to discretize this flow to track it well. In this section, we show that Algorithm 3 is such a natural discretization; in other words, building a deep neural network by stacking gradient layers is a discretization procedure of the gradient flow. We refer to \cite{2} for detailed descriptions on this subject, and also refer to \cite{18} for an original method developed by Otto.

6.1 Continuity Equation and Discretization

We denote by $\mathcal{P}$ the set of probability measures on $\mathbb{R}^v$. For $\mu \in \mathcal{P}$, let $\{\phi_t\}_{t\in[0, \delta]}$ be a curve in $L^2(\mu)$ that solves the following ordinary differential equation: for an $L^2(\mu_{\#} \mu)-$integrable vector field $\nu_t$ on $\mathbb{R}^v$,

$$
\phi_0 = id, \quad \frac{d}{dt} \phi_t(x) = \nu_t(\phi_t(x)) \quad \forall x \in \mathbb{R}^v.
$$

We set $\nu_t = \phi_{\#} \mu$. For simplicity, we assume that $\nu_t$ have smooth density functions $d\nu_t/dx$ with respect to $t \in [0, \delta]$. We denote by $C^\infty_c(\mathbb{R}^v)$ the set of $C^\infty$-functions with compact support in $\mathbb{R}^v$. Using integration by parts, we have that $\forall f \in C^\infty_c(\mathbb{R}^v)$,

$$
\int_{\mathbb{R}^v} f(x) \frac{d}{dt} \nu_t(x) dx = -\int_{\mathbb{R}^v} f(x) \nabla \cdot (\nu_t \nu_t) dx,
$$

where $\nabla \cdot$ is the divergence operator in the weak sense. That is, this equation implies the equality between $d\nu_t/dt$ and $-\nabla \cdot (\nu_t \nu_t)$ as the distribution on $C^\infty_c(\mathbb{R}^v)$. In general, for a Borel family of probability measures $\nu_t$, on $\mathbb{R}^v$, defined for $t$ in the open interval $I \subset \mathbb{R}$ and for a Borel vector field $\nu : (x, t) \to \nu_t(x) \in \mathbb{R}^v$, the following distribution equation in $\mathbb{R}^v \times I$,

$$
\frac{d}{dt} \nu_t + \nabla \cdot (\nu_t \nu_t) = 0
$$

(3)

is called the continuity equation. In other words,

$$
\int_{\mathbb{R}^v} \int_{\mathbb{R}} (\partial_t f(x, t) + \nabla_x f(x, t) \nu_t) d\nu_t dt = 0.
$$

for $\forall f \in C^\infty_c(\mathbb{R}^v \times I)$. Conversely, a narrowly continuous family of probability measures $\nu_t$ solving equation (3) can be obtained by transport map $\phi_t$ satisfying $\phi_t(\phi_0(x)) = \nu_t(x)$ \cite{2}. Thus, equation (3) indicates that the vector field $\nu_t$ drifts the probability measures $\nu_t$.

Here, we focus on curves in the subset $\mathcal{P}_d \subset \mathcal{P}$ composed of probability measures with finite second moment. Noting that there is freedom in the choice of $\nu_t$
modulo divergence-free vector fields \( w \in L^2(\nu_t) \) (i.e., \( \nabla \cdot (w \nu_t) = 0 \)), it is natural to consider the equivalence class of \( v \in L^2(\nu_t) \) modulo divergence-free vector fields. Moreover, there exists a unique \( \Pi(v) \) that attains the minimum \( L^2(\nu_t) \)-norm in this class: \( \Pi(v) = \arg \min_{w \in L^2(\nu_t)} \{ \| w \|_{L^2(\nu_t)} \mid \nabla \cdot ((v - w) \nu_t) = 0 \} \). Thus, we here introduce the definitions of the tangent space at \( \mu \in P_2 \) as follows:

\[
T_\mu P_2 \overset{\text{def}}{=} \{ \Pi(v) \mid v \in L^2(\mu) \}.
\]

The continuity equation \( \delta \) and the tangent space \( \delta \) well characterize the class of absolutely continuous curves in \( P_2 \) in the following sense. For arbitrary narrowly continuous curve \( \nu_t \) (\( t \in \delta \)) in \( P_2 \), the absolute continuity of the curve \( \nu_t \) and satisfying the continuity equation \( \delta \) for some Borel vector field \( v_t \) are equivalent; moreover, we can take \( v_t \) from \( T_{\nu_t} P_2 \) uniquely for almost everywhere \( t \in \delta \) to satisfy the equation \( \delta \).

The following proposition shows how a perturbation using \( v_t \) can discretize an absolutely continuous curve \( \nu_t \) and how \( v_t \) locally approximates optimal transport maps. We denote the 2-Wasserstein distance by \( W_2 \).

**Proposition 3** \( \mathbf{(2)} \). Let \( \nu_t : \delta \to P_2 \) be an absolutely continuous curve satisfying the continuity equation with a Borel vector field \( v_t \) that is contained in \( T_{\nu_t} P_2 \) almost everywhere \( t \in \delta \). Then, for almost everywhere \( t \in \delta \) the following property holds:

\[
\lim_{\delta \to 0} \frac{W_2(\nu_{t+\delta}, (id + \delta v_t) \nu_t)}{|\delta|} = 0.
\]

In particular, for almost everywhere \( t \in \delta \) such that \( \nu_t \) is absolutely continuous with respect to the Lebesgue measure, we have

\[
\lim_{\delta \to 0} \frac{1}{\delta} (t^{\nu_t}_{t+\delta} - id) = v_t \quad \text{in} \quad L^2(\nu_t),
\]

where \( t^{\nu_t}_{t+\delta} \) is the unique optimal transport map between \( \nu_t \) and \( \nu_{t+\delta} \).

This proposition suggests the update \( \mu^+ \leftarrow (id + v_t) \mu \) for discretizing an absolutely continuous curve in \( P_2 \). Note that when \( \mu = \phi_t \nu, (\nu \in P_2, \phi \in L^2(\nu)) \), the corresponding map to \( \mu^+ \) is obtained by \( \phi^+_t \nu = \mu^+ \) where \( \phi^+ \) is a composition as follows:

\[
\phi^+ \leftarrow (id + \xi) \circ \phi = \phi + \xi \circ \phi.
\]

So far, we have introduced the characterization of absolutely continuous curves and a method of their discretization. We notice that the above update resembles the update of Algorithm \( \mathbf{3} \). Indeed, we show that the functional gradient method is nothing but a discretization method of the gradient flow derived by the functional gradient \( \nabla_{\phi} \mathcal{L}(\phi) \).

### 6.2 Discretization of Gradient Flow

We here introduce the gradient flow, which is one of the most straightforward ways to understand Algorithm \( \mathbf{3} \). We have explained that an absolutely continuous curve \( \{\nu_t\}_{t \in \delta} \) in \( P_2 \) is well characterized by the continuity equation \( \delta \) and we have seen that \( \{\nu_t\}_{t \in \delta} \) in \( \delta \) corresponds to the notion of the velocity field induced by the curve. Such a velocity points in the direction of the particle flow. Moreover, the functional gradient \( \nabla_{\phi} \mathcal{L}(\phi) \) points in an opposite direction to reduce the objective \( \mathcal{L} \) at each particle. Thus, these two vector fields exist in the same space and it is natural to consider the following equation:

\[
v_t = -\nabla_{\phi} \mathcal{L}(\phi_t).
\]

This equation for an absolutely continuous curve is called the gradient flow \( \mathbf{2} \) and a curve satisfying this will reduce the objective \( \mathcal{L} \). Indeed, we can find by the chain rule such a curve \( \{\nu_t = \phi_{t+\delta} \mu_t\}_{t \in \delta} \) that also satisfies the following:

\[
\frac{d}{dt} \mathcal{L}(\phi_t) = -\|\nabla_{\phi} \mathcal{L}(\phi_t)\|_{L^2(\nu_t)}^2.
\]

Recalling that \( v_t \) can be discretized well by \( v_{t+\delta} \sim (id - \delta \nabla_{\phi} \mathcal{L}(\phi_t)) \mu_{t+\delta} \), we notice that Algorithm \( \mathbf{3} \) is a discretization method of the gradient flow \( \mathbf{6} \). In other words, building deep neural networks by stacking gradient layers is such a discretization procedure.

### 7 Experiments

In this section, we show the powerful optimization ability of the gradient layer method empirically on training WGANs. Our implementation is done using Theano \( \mathbf{22} \). We first used three toy datasets: swiss roll, 8-gaussian, and 25-gaussian datasets (see Figure \( \mathbf{2} \)) to confirm the convergence behavior of the gradient layer. The sizes of toy datasets are 500, 500, and 1000, respectively. We next used the CIFAR-10 dataset \( \mathbf{13} \) that consists of 50,000 training color images of size \( 32 \times 32 \) with 10-classes and reported inception scores \( \mathbf{20} \), which is one of the conventional scores commonly used to measure the quality of generated samples.

**Toy datasets** We ran Algorithm \( \mathbf{1} \) without pre-training of generators (i.e., \( q = id \)) on toy datasets from Gaussian noise distributions with the standard deviation 0.5. We used four-layer neural networks for the critics where the dimension of hidden layers were set to 128 for swiss roll and 8-gaussian datasets and 512 for 25-gaussian dataset. We adopted one-sided penalty with regularization parameter \( \lambda = 10 \). The output of generator was activated by \( \tanh \). We
Gradient Layer: Enhancing the Convergence of Adversarial Training for Generative Models

Figure 2: Generated samples by Algorithm 1 on 8-gaussian dataset for 0, 25, 50, and 100 generator iterations (from left to right) and training data (rightmost).

Figure 3: Left: Inception scores obtained by WGAN-GP, Right: Inception scores obtained by Algorithm 1 starting from the result of WGAN-GP.

We used ADAM for training critics with parameters $\alpha = 10^{-4}, \beta_1 = 0.5, \beta_2 = 0.9$, minibatch size 50. When we run Algorithm 1, gradient layers are stacked below tanh. The learning rates were set to $\eta = 0.1$. The number of inner iterations $T_0$ for training the critics was set to $\text{datasize} \times 5$. Figure 2 shows the results for toy datasets for running $T = 100$ iterations of generators. Although we ran the algorithm without pre-training the generators, we obtained better results only for a few iterations. This is surprising, because these toy datasets are difficult to learn and fail to converge in the standard GANs and WGANs. Whereas improved variants of these models overcome this difficulty, they usually require more than 1,000 iterations to converge.

CIFAR-10 dataset We first trained WGAN-GP with a two-sided penalty ($\lambda = 10$) on the CIFAR-10 dataset. We used DCGAN for both the critic and the generator. The batch normalization [10] was used only for the generator. The critic and the generator were trained by using ADAM with $\alpha = 10^{-4}, \beta_1 = 0.5, \beta_2 = 0.9$, and minibatch size 64. The number of inner iterations for training the critic was 5 and we ran ADAM for $10^5$-iterations. The left side of Figure 3 shows the inception scores obtained by WGAN-GP. It seems that the learning procedure is slowed down and gets stuck in a final training phase. The final inception score is 6.32. We next ran Algorithm 1 starting from the result of WGAN-GP. The critic was trained by ADAM with the same parameters, except for $\alpha = 5 \times 10^{-5}$ and $T_0 = \text{datasize} \times 1$. The number of gradient layers was 20 with the learning rate $\eta = 0.5$. The right side of Figure 3 shows the inception scores obtained by Algorithm 1 and the generated samples are listed in Figure 1. We observed a rapid increase in the inception score, which was improved to 6.80.

8 Conclusion

We have proposed a gradient layer that enhances the convergence speed of adversarial training. Because this layer is based on the perspective of infinite-dimensional optimization, it can avoid local optima induced by non-convexity and parameterization. We have also provided two perspectives of the gradient layer: (i) the functional gradient method and (ii) the discretization procedure of the gradient flow. We have proven the fast convergence of the gradient layer by utilizing this perspective, and experimental results have empirically shown its reliable performance.
Acknowledgements

This work was partially supported by MEXT KAKENHI (25730013, 25120012, 26280009, 15H01678 and 15H05707), JST-PRESTO (JPMJPR14E4), and JST-CREST (JPMJCR14D7, JPMJCR1304).

References

[1] Luigi Ambrosio. Lecture notes on optimal transport problems. LECTURE NOTES IN MATHEMATICS-SPRINGER VERLAG-, pages 1–52, 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] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. arXiv preprint arXiv:1701.07875, 2017.

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

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

[6] Xi Chen, Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In Advances in Neural Information Processing Systems 29, pages 2172–2180. 2016.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems 27. 2014.

[8] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. arXiv preprint arXiv:1704.00028, 2017.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[10] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International Conference on Machine Learning, pages 448–456, 2015.

[11] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), 2014.

[12] Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. How to train your dragan. arXiv preprint arXiv:1705.07215, 2017.

[13] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

[14] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. arXiv preprint arXiv:1512.09300, 2015.

[15] David G Luenberger. Optimization by vector space methods. John Wiley & Sons, 1969.

[16] Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. Econometrica, 70(2):583–601, 2002.

[17] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In Advances in Neural Information Processing Systems 29, pages 271–279. 2016.

[18] Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. 2001.

[19] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434, 2015.

[20] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In Advances in Neural Information Processing Systems, pages 2234–2242, 2016.

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

[22] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. arXiv preprint arXiv:1605.02688, 2016.
[23] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning, 4(2):26–31, 2012.

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

[25] Dilin Wang and Qiang Liu. Learning to draw samples: With application to amortized mle for generative adversarial learning. arXiv preprint arXiv:1611.01722, 2016.

[26] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. arXiv:1612.03242, 2016.
Appendix

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 4. Note that we always use latest parameters of \( f, g \) for gradient layers in Algorithm 4. When gradient layers are inserted in the middle of the generator: \( g_1 \circ \phi \circ g_2 \), we can apply Algorithm 4 by setting \( \mu_n \leftarrow g_2 \mu_n, g \leftarrow g_1 \). 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 Algorithm 5.

Algorithm 4 Assisting WGAN-GP

Input: The base distribution \( \mu_n \), the minibatch size \( b \), the number of iterations \( T \), the initial parameters \( \tau_0 \) and \( \theta_0 \) of the critic and the generator, the number of iterations \( T_0 \) for the critic, the regularization parameter \( c \), learning rate \( \eta \) 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
    \( \{x_i\}_{i=1}^b \sim \mu_D, \{z_i\}_{i=1}^b \sim \mu_n, \{e_i\}_{i=1}^b \sim U[0, 1]^b \)
    \( G^{\tau_k, \theta_k} \) is applied \( l \) times.
    \( \{z_i\}_{i=1}^b \leftarrow \{g_{\theta_k} \circ G_{\eta}^{\tau_k, \theta_k} \circ \cdots \circ G_{\eta}^{\tau_k, \theta_k}(z_i)\}_{i=1}^b \)
    \( \{\tilde{x}_i\}_{i=1}^b \leftarrow \{\epsilon_i x_i + (1 - \epsilon_i) z_i\}_{i=1}^b \)
    \( v = \nabla_{\theta} \frac{1}{b} \sum_{i=1}^b [f_r(\tilde{x}_i) - f_r(x_i) + \lambda R_{f_r}(\tilde{x}_i)] \)
    \( \tau \leftarrow A(\tau, v) \)
  end for
  \( \tau_{k+1} \leftarrow \tau \)
  \( \{z_i\}_{i=1}^b \sim \mu_n \)
  \( G^{\tau_{k+1}, \theta_k} \) is applied \( l \) times.
  \( \{z_i\}_{i=1}^b \leftarrow \{G_{\eta}^{\tau_{k+1}, \theta_k} \circ \cdots \circ G_{\eta}^{\tau_{k+1}, \theta_k}(z_i)\}_{i=1}^b \)
  \( v \leftarrow -\nabla_{\theta} \frac{1}{b} \sum_{i=1}^b f_{\tau_{k+1}}(g_{\theta_k}(z_i)) \)
  \( \theta_{k+1} \leftarrow A(\theta_k, v) \)
end for
Return \( \tau_T, \theta_T \).

Algorithm 5 Data Generation for Algorithm 4

Input: the seed drawn from the base measure \( z \sim \mu_n \), the parameter \( \tau \) and \( \theta \) of the critic and the generator, the learning rate \( \eta \), 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 4 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 \( \mu_n \) 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 \( p \)-th moment on \( \mathcal{X} \subset \mathbb{R}^c \). For \( \mu, \nu \in \mathcal{P}_p \) a probability measure \( \gamma \) on \( \mathcal{X} \times \mathcal{X} \) satisfying \( \pi^i_\mu \gamma = \mu \) and \( \pi^i_\nu \gamma = \nu \) is called a plan (coupling), where \( \pi^i \) 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 \( \mu \) and \( \nu \). We now introduce Kantorovich’s formulation of the \( p \)-Wasserstein
distance $W_p$ for $p \geq 1$. 

$$W_p^p(\mu, \nu) = \min_{\gamma \in \Gamma(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} \|x - y\|^p d\gamma(x, y) \quad (7)$$

When $p = 1$ and $\mu, \nu$ 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. [24, 2]) and we denote by $\Gamma$ the set of optimal plans. Prior to this formulation, the optimal transport problem in Monge’s formulation was proposed.

$$\inf_{\phi \in \mathcal{M}(\mu, \nu)} \int_{\mathcal{X}} \|x - \phi(x)\|_2^p d\mu(x), \quad (8)$$

where the infimum is taken over all transport maps $\phi : \mathcal{X} \to \mathcal{X}$ from $\mu$ to $\nu$, i.e., $\phi_* \mu = \nu$. Because a transport map $\phi$ gives a plan $\gamma = (id \times \phi)_* \mu$, we can easily find $\mathcal{L} \leq \mathcal{K}$. In general, an optimal transport map that solves the problem $\mathcal{K}$ does not always exist unlike Kantrovich problem $\mathcal{L}$. However, in the case where $p = 1$, $\mathcal{X} = \mathbb{R}^n$, and $\mu$ is absolutely continuous with respect to the Lebesgue measure, the existence of optimal transport maps is guaranteed [4, 5], and it is extended to more general integrand (see [2]). Moreover, this optimal transport map also solves Kantrovich problem $\mathcal{L}$, 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 [21], see also [1]). Let $\mathcal{X}$ be a compact convex subset in $\mathbb{R}^n$ and assume that $\mu$ is absolutely continuous with respect to Lebesgue measure. Then, there exists an optimal transport map $\phi$ from $\mu$ to $\nu$ for the problem $\mathcal{K}$ with $p = 1$. Moreover, if $\nu$ is also absolutely continuous with respect to Lebesgue measure, we can choose $\psi$ so that $\psi^{-1}$ is well defined $\mu_0$-a.e., and $\phi_* \nu = \mu$.

Under the same assumption in Proposition A, it is known that two distances $\mathcal{K}$ and $\mathcal{L}$ coincide [1], that is, the Kantrovich problem $\mathcal{L}$ is solved by an optimal transport map.

### C Proofs

We here the give proof of Proposition [1]

**Proof of Proposition [7]** Note that $\mathcal{L}(\psi) = \hat{\mathcal{L}}(f_{\phi}^*, \psi)$. For $\psi \in B_r^\infty(\phi)$, we divide $\mathcal{L}(\psi)$ into two terms as follows.

$$\mathcal{L}(\psi) = (\hat{\mathcal{L}}(f_{\phi}^*, \psi) - \hat{\mathcal{L}}(f_{\phi}^*, \psi)) + \hat{\mathcal{L}}(f_{\phi}^*, \psi). \quad (9)$$

We first bound the first term in (9) by $L$-smoothness of $\hat{\mathcal{L}}(f_{\phi}^*, \psi)$ with respect to $\psi'$ at $\psi$ in $B_r^\infty(\psi)$.

$$\left| \hat{\mathcal{L}}(f_{\phi}^*, \psi) - (\hat{\mathcal{L}}(f_{\phi}^*, \psi) + (\nabla_{\psi'} \hat{\mathcal{L}}(f_{\phi}^*, \psi)_{\psi' = \psi} \phi' - \psi_2 L^2(\mu_\psi)) \right| \leq \frac{L^2}{2} \|\phi - \psi\|_{L^2(\mu_\psi)}. \quad (10)$$

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

$$\left| \hat{\mathcal{L}}(f_{\phi}^*, \psi) - \hat{\mathcal{L}}(f_{\phi}^*, \psi') \right| \leq \frac{L^2}{2} \|\phi - \psi\|_{L^2(\mu_\psi)}. \quad (11)$$

We next bound $\hat{\mathcal{L}}(f_{\phi}^*, \psi)$ in (9). We remember that

$$\hat{\mathcal{L}}(f_{\phi}^*, \psi) = \mathbb{E}_{x \sim \mu_{\phi}}[f_{\phi}^*(x)] - \mathbb{E}_{x \sim \mu_{\phi}}[f_{\phi}^* \circ \psi(x)] - \lambda Rf_{\phi}^*. \quad (11)$$

By $L$-smoothness of $f_{\phi}^*$, it follows that

$$\left| f_{\phi}^*(\psi(x)) - (f_{\phi}^*(\phi(x))) + (\nabla_z f_{\phi}^*(z)|_{z = \phi(x)}, \psi(x) - \phi(x))_2 \right| \leq \frac{L^2}{2} \|\psi(x) - \phi(x)\|_2^2. \quad (12)$$

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

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


We substitute this inequality into (11), we have
\[ \hat{L}(f^*_\phi, \psi) \leq E_{x \sim \mu_D} [f^*_\phi(x)] + \frac{L}{2} \| \psi - \phi \|^2_{L^2(\mu_\phi)} - (E_{\mu_\phi} [f^*_\phi(\phi(x))] + \langle \nabla_x f^*_\phi \circ \phi, \psi - \phi \rangle_{L^2(\mu_\phi)}) - \lambda R_{f^*_\phi} \]
\[ = \hat{L}(f^*_\phi, \phi) - \langle \nabla_x f^*_\phi \circ \phi, \psi - \phi \rangle_{L^2(\mu_\phi)} + \frac{L}{2} \| \psi - \phi \|^2_{L^2(\mu_\phi)} \]
\[ = \mathcal{L}(\phi) + \langle \nabla_{\phi} \mathcal{L}(\phi), \psi - \phi \rangle_{L^2(\mu_\phi)} + \frac{L}{2} \| \psi - \phi \|^2_{L^2(\mu_\phi)}, \tag{12} \]
and the opposite inequality
\[ \hat{L}(f^*_\phi, \psi) \geq \mathcal{L}(\phi) + \langle \nabla_{\phi} \mathcal{L}(\phi), \psi - \phi \rangle_{L^2(\mu_\phi)} - \frac{L}{2} \| \psi - \phi \|^2_{L^2(\mu_\phi)}, \tag{13} \]
where we used \( \nabla_{\phi} \mathcal{L}(\phi) = -\nabla_x f^*_\phi(z)|_{z=\phi(\cdot)}. \) By combining (9), (10), and (12), we have
\[ \mathcal{L}(\psi) \leq \mathcal{L}(\phi) + \langle \nabla_{\phi} \mathcal{L}(\phi), \psi - \phi \rangle_{L^2(\mu_\phi)} + L \| \phi - \psi \|^2_{L^2(\mu_\phi)}. \]
Moreover, since \( \hat{L}(f^*_\phi, \psi) - \hat{L}(f^*_\psi, \psi) \geq 0 \) in (9), we have \( \mathcal{L}(\psi) \geq \hat{L}(f^*_\psi, \psi). \) Therefore, we get the opposite inequality by (13)
\[ \mathcal{L}(\psi) \geq \mathcal{L}(\phi) + \langle \nabla_{\phi} \mathcal{L}(\phi), \psi - \phi \rangle_{L^2(\mu_\phi)} - \frac{L}{2} \| \phi - \psi \|^2_{L^2(\mu_\phi)}. \]
This finishes the proof. \[ \square \]

We next provide the proof of Theorem 1.

Proof of Theorem 1\[ \square \] Noting that \( \| \eta \nabla_{\phi_k} \mathcal{L}(\phi_k) \|_{\infty} \leq r \) and Lipschitz smoothness of \( \mathcal{L}, \) we have
\[ \mathcal{L}(\phi_{k+1}) \leq \mathcal{L}(\phi_k) \leq \mathcal{L}(\phi_k) - \eta \| \nabla_{\phi} \mathcal{L}(\phi_k) \|^2_{L^2(\mu_\phi)} + \frac{\eta^2 L}{2} \| \nabla_{\phi} \mathcal{L}(\phi_k) \|^2_{L^2(\mu_\phi)} \]
\[ = \mathcal{L}(\phi_k) - \eta(1 - \eta L/2) \| \nabla_{\phi} \mathcal{L}(\phi_k) \|^2_{L^2(\mu_\phi)}. \]
Since \( \eta \leq 1/L, \) we have \( \mathcal{L}(\phi_{k+1}) \leq \mathcal{L}(\phi_k) - \frac{\eta}{2} \| \nabla_{\phi} \mathcal{L}(\phi_k) \|^2_{L^2(\mu_\phi)}. \) Summing up over \( k \in \{0, \ldots, T-1\} \) and dividing by \( T \) we obtain
\[ \frac{1}{T} \sum_{k=0}^{T-1} \| \nabla_{\phi} \mathcal{L}(\phi_k) \|^2_{L^2(\mu_\phi)} \leq \frac{2}{\eta T} (\mathcal{L}(\phi_0) - \mathcal{L}(\phi_T)). \]
This inequality finishes the proof of the theorem. \[ \square \]

Proof of Proposition 2\[ \square \]. By Proposition 1\[ \square \] there exists an optimal transport \( \psi \) from \( \mu_\phi \) to \( \mu_D \) and an optimal plan is given by \( \gamma = (id \times \psi)_{\mu_\phi} \). We set \( \psi_t = (1-t)id + t\psi \) and \( \mu_t \equiv \psi_{t}\mu_\phi. \) Because \( (\psi_s, \psi_t)_{t\mu_\phi} \) gives a plan between \( \mu_\phi \) and \( \mu_D \), we have
\[ W_1(\mu_s, \mu_t) \leq \int_{X \times X} |x - y|_2 d(\psi_s, \psi_t)_{\mu_\phi} \]
\[ = \int_X |\psi_s(x) - \psi_t(x)|_2 d\mu_\phi \]
\[ = (t-s) \int_X \| x - \psi(x) \|_2 d\mu_\phi = (t-s)W_1(\mu_\phi, \mu_D). \tag{14} \]
We next prove the opposite inequality. Noting that \( (id, \psi_s)_{t\mu_\phi} \) is a plan from \( \mu_\phi \) to \( \mu_s \) and \( (\psi_t, \psi_t)_{t\mu_\phi} \) is a plan from \( \mu_t \) to \( \mu_D, \) we have the following two inequalities
\[ W_1(\mu_g, \mu_s) \leq \int_{X \times X} |x - y|_2 d(id, \psi_s)_{\mu_\phi} = \int_X \| x - \psi_s(x) \|_2 d\mu_s = sW_1(\mu_\phi, \mu_D), \]
\[ W_1(\mu_s, \mu_D) \leq \int_{X \times X} |x - y|_2 d(\psi_t, \psi_t)_{\mu_\phi} = \int_X \| \psi_t(x) - \psi_t(x) \|_2 d\mu_\phi = (1-t)W_1(\mu_\phi, \mu_D). \]
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 (14), 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 4.

![Figure 4](image)

Figure 4: Random samples drawn from the generator trained by WGAN-GP (left) and the gradient layer (right).