Supplementary: Do GANs always have Nash equilibria?

Farzan Farnia [1] **Asuman Ozdaglar** [1]

1. Numerical Results for Section 2

Here, we provide the complete numerical results for the experiments discussed in Section 2 of the main text. Regarding the plots shown in Section 2 for the SN-GAN implementation, here we present the same plots for the Wasserstein GAN with weight clipping (WGAN-WC) and with gradient penalty (WGAN-GP) problems. Figures 1-4 repeat the experiments of Figures 1,2 in the main text for the WGAN-WC and WGAN-GP problems. These plots suggest that a similar result also holds for the WGAN-WC and WGAN-GP problems, where the objective and the generated samples' quality were decreasing during the generator optimization. For a larger set of generated samples in the main text's Figures 1,2 and Figures 1-4, we refer the readers to Figures 5-10.

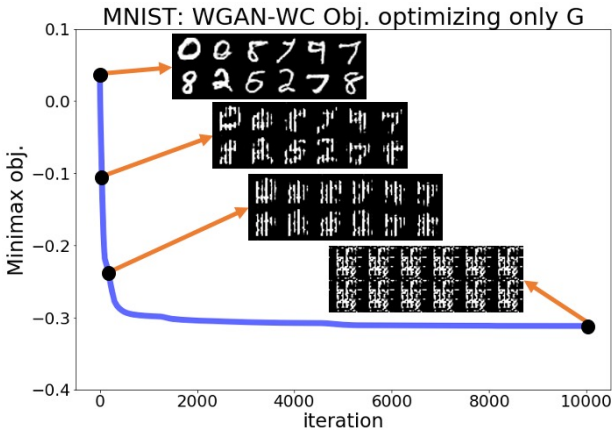

Figure 1. Optimizing the trained generator of WGAN-WC with a fixed discriminator on MNIST data. The GAN's objective and samples' quality were decreasing over the optimization.

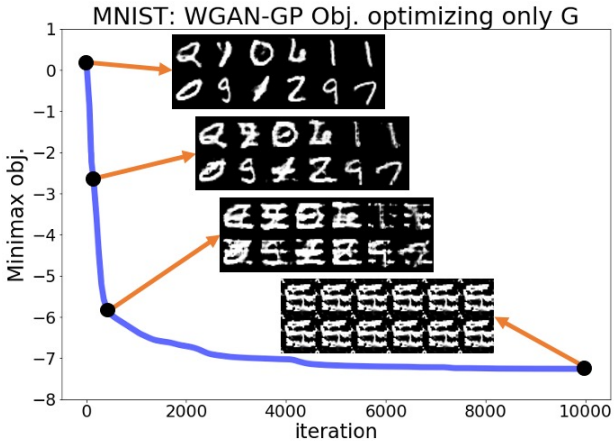

Figure 2. Repeating the experiment of Figure 1 for WGAN-GP.

2. Numerical Results for Section 6

Here, we present the complete numerical results for the experiments of Section 6 in the main text. Figures 11-14 demonstrate the results of the main text's Figures 4,5 for the WGAN-WC and WGAN-GP problems. Here, except the WGAN-GP experiment on the CelebA dataset, we observed that the objective and the generated samples' quality did not significantly decrease over the generator optimization. Even for the WGAN-GP experiment on the CelebA data, we observed that the objective value decreased three times less than in minimizing the original objective rather than the proximal objective. These experiments suggest that the Wasserstein and Lipschitz GAN problems can converge to local proximal equilibrium solutions. We also show a larger group of generated samples at the beginning and final iterations of Figures 4,5 in the main text and Figures 11-14 in Figures 15-20.

For the proximal training experiments, Figures 21-25 show the samples generated by the SN-GAN and WGAN-WC proximally trained on CIFAR-10 and CelebA data with the results for the baseline regular training on the top of the figure and the results for proximal training on the bottom. We observed a somewhat improved quality achieved by proximal training, which was further supported by the inception scores for the CIFAR-10 experiments reported in the main text.

[1]Laboratory for Information & Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, USA. Correspondence to: Farzan Farnia <farnia@mit.edu>, Asuman Ozdaglar <asuman@mit.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

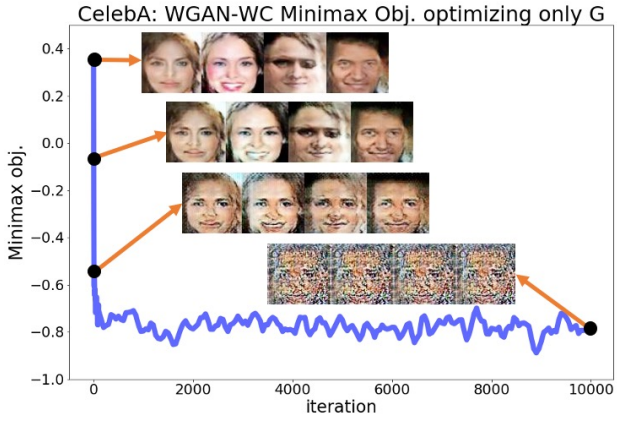

Figure 3. Repeating the experiment of Figure 1 for the CelebA data.

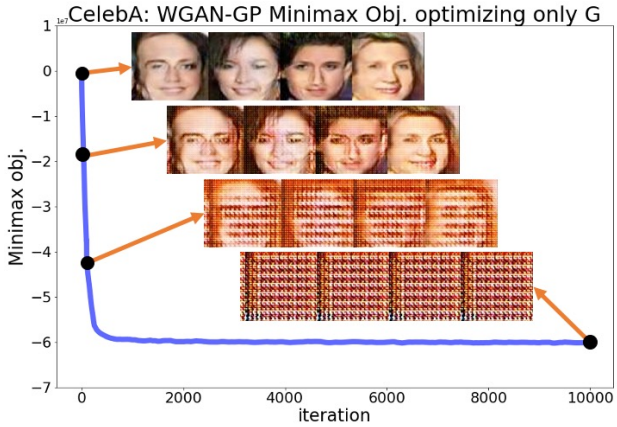

Figure 4. Repeating the experiment of Figure 3 for WGAN-GP.

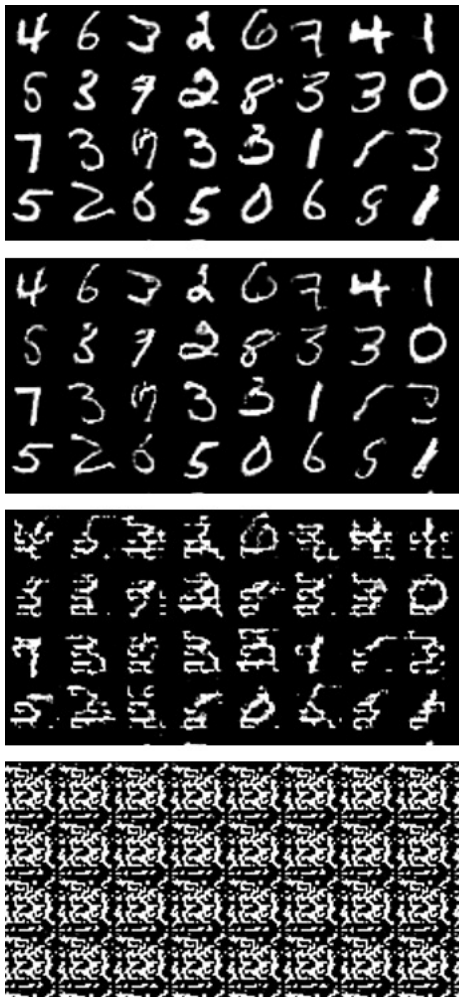

Figure 5. SN-GAN's generated samples at the iterations marked in the main text's Figure 1.

3. Proofs

3.1. Proof of Proposition 1

Proposition 1. *Assume generator function $G^* \in \mathcal{G}$ results in the distribution of data $\mathbf{X}$, i.e. we have $P_{G^*(\mathbf{Z})} = P_{\mathbf{X}}$. Then, for each of the GAN problems discussed in the main text there exists a constant discriminator function D_{constant} which together with G^* results in a Nash equilibrium and satisfies the following for every $G \in \mathcal{G}$ and $D \in \mathcal{D}$:*

$$V(G^*, D) \leq V(G^*, D_{\text{constant}}) \leq V(G, D_{\text{constant}}).$$

Proof. **Proof for f-GANs:**

Consider the following f-GAN minimax problem corresponding to the convex function f:

$$\min_{G \in \mathcal{G}} \max_{D} \mathbb{E}\big[D(\mathbf{X})\big] - \mathbb{E}\big[f^*(D(G(\mathbf{Z})))\big]. \tag{1}$$

Due to the realizability assumption, given $G^* \in \mathcal{G}$ we assume that the data distribution and the generative model are identical, i.e., $P_{\mathbf{X}} = P_{G^*(\mathbf{Z})}$. Then, the minimax objective for G^* reduces to

$$\mathbb{E}_{P_{\mathbf{X}}}\big[\, D(\mathbf{X}) - f^*(D(\mathbf{X}))\,\big]. \tag{2}$$

The above objective decouples across $\mathbf{X}$ outcomes. As a result, the maximizing discriminator $D^*(\mathbf{x}) = f'(1)$ will be a constant function where the constant value $f'(1)$ follows from the optimization problem:

$$f'(1) = \operatorname*{argmax}_{u \in \mathbb{R}} u - f^*(u). \tag{3}$$

Note that the objective $u - f^*(u)$ is a concave function of u whose derivative is zero at ${f^*}'^{-1}(1) = f'(1)$, because the Fenchel-conjugate of a convex f satisfies ${f^*}'^{-1} = f'$.

So far we have proved that the constant function $D_{\text{constant}}(\mathbf{x}) = f'(1)$ provides the optimal discriminator for generator G^*. Therefore, for every discriminator D we

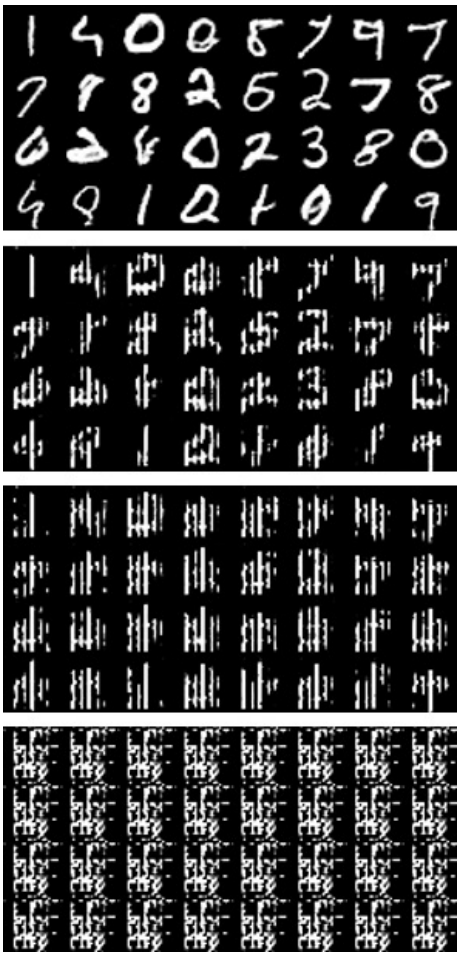

Figure 6. WGAN-WC's generated samples at the iterations marked in Figure 1.

Figure 7. WGAN-GP's generated samples at the iterations marked in Figure 2.

have

$$V(G^*, D) \leq V(G^*, D_{\text{constant}}), \tag{4}$$

where $V(G, D)$ denotes the f-GAN's minimax objective. Moreover, note that for a constant D the value of the minimax objective does not change with generator G. As a result, for every G

$$V(G, D_{\text{constant}}) = V(G^*, D_{\text{constant}}). \tag{5}$$

Then, (4) and (5) collectively prove that for every G and D we have

$$V(G^*, D) \leq V(G^*, D_{\text{constant}}) \leq V(G, D_{\text{constant}}),$$

which completes the proof for f-GANs.

Proof for Wasserstein GANs:

Consider a general Wasserstein GAN problem with a cost function c satisfying $c(\mathbf{x}, \mathbf{x}) = 0$ for every $\mathbf{x}$. Notice that this property holds for all Wasserstein distance measures corresponding to cost function $\|\mathbf{x}-\mathbf{x}'\|^q$ for $q \geq 1$. The generalized Wasserstein GAN minimax problem is as follows:

$$\min_{G \in \mathcal{G}} \max_{D \ c\text{-concave}} \mathbb{E}[D(\mathbf{X})] - \mathbb{E}\big[D^c(G(\mathbf{Z}))\big]. \tag{6}$$

Due to the realizability assumption, a generator function $G^* \in \mathcal{G}$ results in the data distribution such that $P_{G^*(\mathbf{Z})} = P_{\mathbf{X}}$. Then, the above minimax objective for G^* reduces to

$$\mathbb{E}_{P_{\mathbf{X}}}\big[\, D(\mathbf{X}) - D^c(\mathbf{X}) \,\big]. \tag{7}$$

Since the cost is assumed to take a zero value given identical inputs, we have:

$$\begin{aligned} D^c(\mathbf{x}) :&= \max_{\mathbf{x}'} \ D(\mathbf{x}') - c(\mathbf{x}, \mathbf{x}') \\ &\geq D(\mathbf{x}) - c(\mathbf{x}, \mathbf{x}) \\ &= D(\mathbf{x}). \end{aligned}$$

As a result, $D(\mathbf{x}) - D^c(\mathbf{x}) \leq 0$ holds for every $\mathbf{x}$. Hence, the objective in (7) will be non-positive and takes its maximum

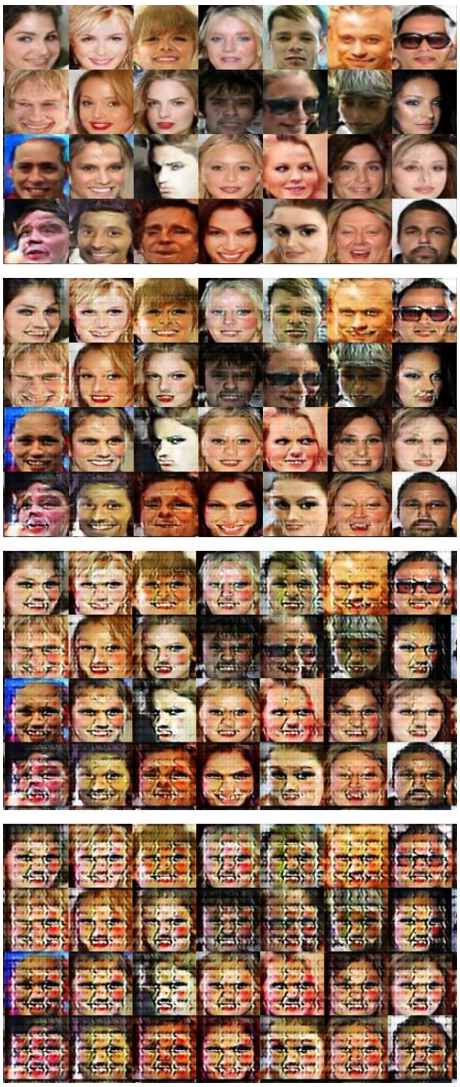

Figure 8. SN-GAN's generated samples at the iterations marked in the main text's Figure 2.

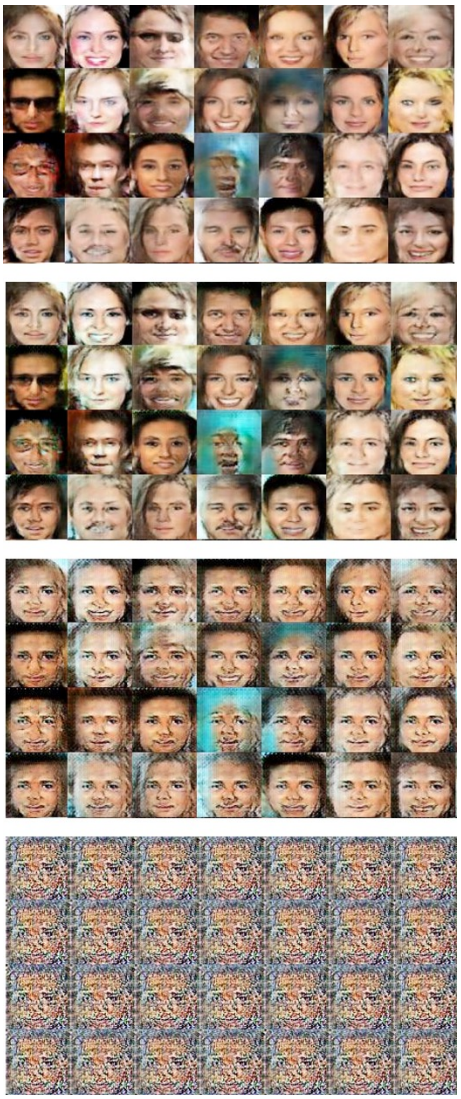

Figure 9. WGAN-WC's generated samples at the iterations marked in Figure 3.

zero value for any constant function D_{constant}, which by definition satisfies c-concavity. Therefore, letting $V(G, D)$ denote the GAN minimax objective, for every D we have

$$V(G^*, D) \leq V(G^*, D_{\text{constant}}). \tag{8}$$

We also know that for a constant discriminator D_{constant} the value of the minimax objective is independent from the generator function. Therefore, for every G we have

$$V(G^*, D_{\text{constant}}) = V(G, D_{\text{constant}}). \tag{9}$$

As a result, (8) and (9) together show that for every G and D

$$V(G^*, D) \leq V(G^*, D_{\text{constant}}) \leq V(G, D_{\text{constant}}), \tag{10}$$

which makes the proof complete for Wasserstein GANs. □

3.2. Proof of Theorem 1 & Remark 1

Theorem 1. *Consider a GAN minimax problem for learning a normally distributed $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$ with zero mean and scalar covariance matrix where $\sigma > 1$. In the GAN formulation, we use a linear generator function $G(\mathbf{z}) = \mathbf{W}\mathbf{z} + \mathbf{u}$ where the weight matrix $\mathbf{W}$ is spectrally-regularized to satisfy $\sigma_{\max}(\mathbf{W}) \leq 1$ and $\|\mathbf{u}\|_2 \leq t$ for constant $t > 0$. Suppose that the Gaussian latent vector is normally distributed as $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, I)$ with zero mean and identity covariance matrix. Then,*

- *For the f-GAN problem corresponding to an f with non-decreasing $t^2 f''(t)$ over $t \in (0, +\infty)$ and an unconstrained discriminator D where the dimensions of data $\mathbf{X}, \mathbf{Z}$ are equal, the f-GAN minimax problem has no Nash*

Figure 10. WGAN-GP's generated samples at the iterations marked in Figure 4.

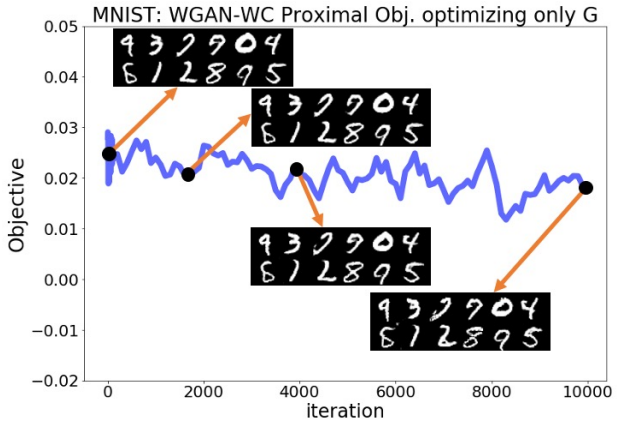

Figure 11. Optimizing the proximal objective for the trained generator in WGAN-WC with a fixed discriminator on MNIST data. The GAN's objective and samples' quality were preserved over the optimization.

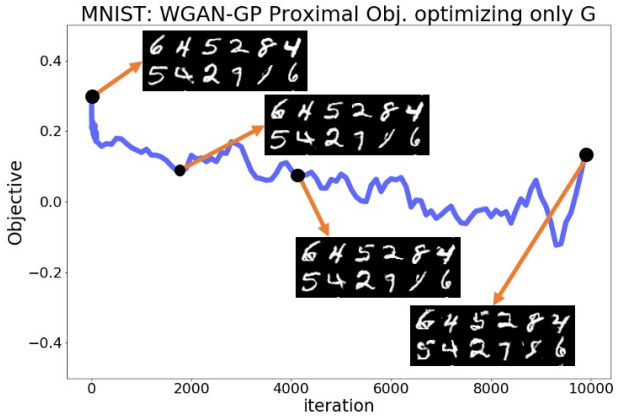

Figure 12. Repeating the experiment of Figure 11 for the WGAN-GP problem.

equilibrium solutions.

- *For the W2GAN problem with discriminator D trained over c-concave functions, where c is the quadratic cost, the W2GAN minimax problem has no Nash equilibrium solutions. Also, given a quadratic discriminator $D(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} + \mathbf{b}^T\mathbf{x}$ parameterized by $A, \mathbf{b}$, the W2GAN problem has no **local** Nash equilibria.*
- *For the WGAN problem with 1-dimensional X, Z and a discriminator D trained over 1-Lipschitz functions, the WGAN minimax problem has no Nash equilibria.*

Proof. **Proof for f-GANs:**

Lemma 1. *Consider two random vectors $\mathbf{X}, \widetilde{\mathbf{X}}$ with probability density functions p, q, respectively. Suppose that p, q are non-zero everywhere. Then, considering the following variational representation of $d_f(P, Q)$,*

$$d_f(P,Q) = \max_D \mathbb{E}[D(\mathbf{X})] - \mathbb{E}[f^*(D(\widetilde{\mathbf{X}}))], \tag{11}$$

the optimal solution D^* will satisfy

$$D^*(\mathbf{x}) = f'\big(\frac{p(\mathbf{x})}{q(\mathbf{x})}\big). \tag{12}$$

Proof. Let us rewrite the f-divergence's variational representation as

$$\begin{aligned} d_f(P,Q) &= \max_D \mathbb{E}[D(\mathbf{X})] - \mathbb{E}[f^*(D(\widetilde{\mathbf{X}}))] \\ &= \max_D \int \big[p(\mathbf{x})D(\mathbf{x}) - q(\mathbf{x})f^*(D(\mathbf{x}))\big]\,\mathrm{d}\mathbf{x} \\ &= \int \max_{D(\mathbf{x})} \big[p(\mathbf{x})D(\mathbf{x}) - q(\mathbf{x})f^*(D(\mathbf{x}))\big]\,\mathrm{d}\mathbf{x} \end{aligned}$$

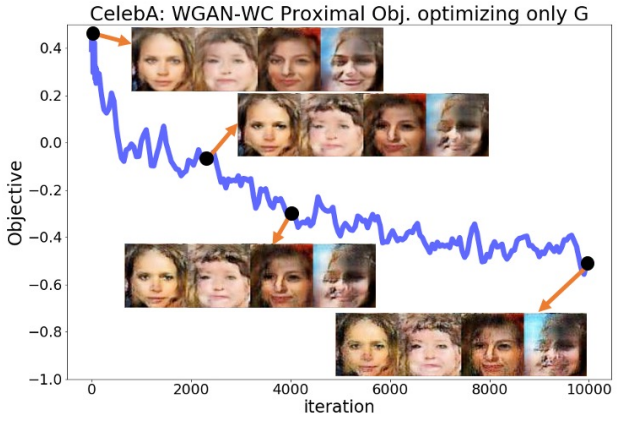

Figure 13. Optimizing the proximal objective for the trained generator in WGAN-WC with a fixed discriminator on CelebA data. The GAN's objective and samples' quality were preserved over the optimization.

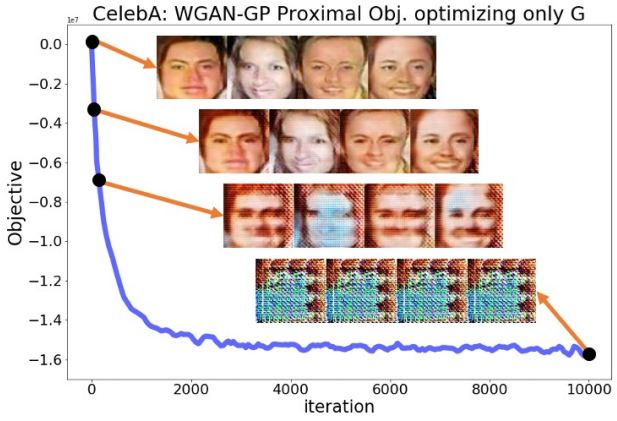

Figure 14. Repeating the experiment of Figure 13 for WGAN-GP. The samples quality and objective value were decreasing for WGAN-GP on CelebA.

where the last equality holds, since the maximization objective decouples across $\mathbf{x}$ values. It can be seen that the inside optimization problem for each $D(\mathbf{x})$ is maximizing a concave objective in which by setting the derivative to zero we obtain

$$f^{*\prime}(D^*(\mathbf{x})) = \frac{p(\mathbf{x})}{q(\mathbf{x})}. \tag{13}$$

As a property of the Fenchel-conjugate of a convex f, we know $f^{*\prime -1} = f^{\prime}$ which combined with the above equation implies that

$$D^*(\mathbf{x}) = f^{\prime}\big(\frac{p(\mathbf{x})}{q(\mathbf{x})}\big). \tag{14}$$

The above result completes Lemma 1's proof. □

Consider the f-GAN problem with the generator function

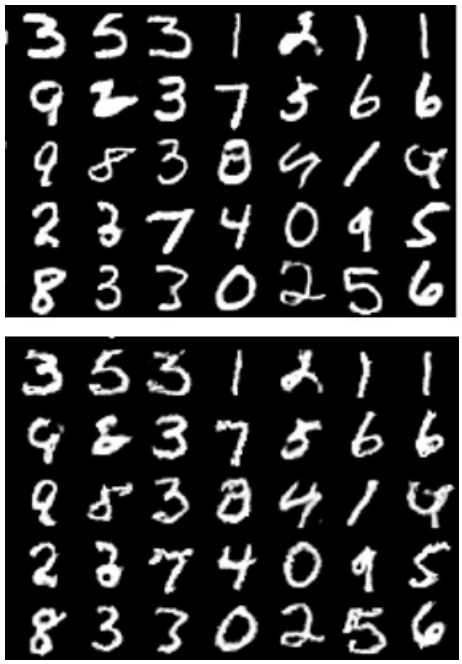

Figure 15. SN-GAN's generated samples at the first and last iterations of the main text's Figure 5.

specified in the theorem:

$$\min_{\mathbf{W},\mathbf{u}:\,\|\mathbf{W}\|_2\le 1}\ \max_{D}\ \mathbb{E}\big[D(\mathbf{X})\big] - \mathbb{E}\big[f^*\big(D(\mathbf{W}\mathbf{Z}+\mathbf{u})\big)\big]. \tag{15}$$

Note that $\mathbf{X}\sim\mathcal{N}(\mathbf{0},\sigma^2 I)$ and $\mathbf{W}\mathbf{Z}+\mathbf{u}\sim\mathcal{N}(\mathbf{u},\mathbf{W}\mathbf{W}^T)$. Notice that if $\mathbf{W}$ was not full-rank, the maximized discriminator objective would be $+\infty$ achieved by a D assigning an infinity value to the points not included in the rank-constrained support set of generator $\mathbf{W}\mathbf{z}+\mathbf{u}$. This will not result in a solution to the f-GAN problem, because we assume that the dimensions of $\mathbf{X}$ and $\mathbf{Z}$ match each other and hence there exists a full-rank $\mathbf{W}$ with a finite maximized objective, i.e. f-divergence value. Therefore, in a Nash equilibrium of the f-GAN problem, the solution $\mathbf{W}$ must be full-rank and invertible.

Lemma 1 results in the following equation for the optimal discriminator $D^*_{\mathbf{W},\mathbf{u}}$ given generator parameters $\mathbf{W},\mathbf{u}$:

$$\begin{aligned}
D^*_{\mathbf{W},\mathbf{u}}(\mathbf{x}) \;&=\; f^{\prime}\Bigg(\frac{\frac{1}{\sqrt{(2\pi\sigma^2)^k}}\exp\big\{-\frac{1}{2\sigma^2}\|\mathbf{x}\|_2^2\big\}}{\frac{1}{\sqrt{(2\pi)^k\det(\mathbf{W}\mathbf{W}^T)}}\exp\big\{-\frac{1}{2}\big\|(\mathbf{W}\mathbf{W}^T)^{-1/2}(\mathbf{x}-\mathbf{u})\big\|_2^2\big\}}\Bigg)\\
&= f^{\prime}\Bigg(\sqrt{\frac{\det(\mathbf{W}\mathbf{W}^T)}{\sigma^{2k}}}\exp\Big\{\frac{1}{2}\mathbf{x}^T((\mathbf{W}\mathbf{W}^T)^{-1}-\sigma^{-2}I)\mathbf{x}\\
&\quad -\mathbf{u}^T(\mathbf{W}\mathbf{W}^T)^{-1}\mathbf{x}+\mathbf{u}^T(\mathbf{W}\mathbf{W}^T)^{-1}\mathbf{u}\Big\}\Bigg).
\end{aligned}$$

Figure 16. WGAN-WC's generated samples at the first and last iterations of Figure 11.

Figure 17. WGAN-GP's generated samples at the first and last iterations of Figure 12.

As a result, the function $f^*(D^*_{\mathbf{W},\mathbf{u}}(\cdot))$ appearing in the f-GAN's minimax objective will be

$$f^*\big(D^*_{\mathbf{W},\mathbf{u}}(\mathbf{x})\big) = f^*\Bigg(f'\Bigg(\sqrt{\frac{\det(\mathbf{W}\mathbf{W}^T)}{\sigma^{2k}}}\exp\Big\{\frac{1}{2}\mathbf{x}^T((\mathbf{W}^T\mathbf{W})^{-1} - \sigma^{-2}I)\mathbf{x} - \mathbf{u}^T(\mathbf{W}\mathbf{W}^T)^{-1}\mathbf{x} + \mathbf{u}^T(\mathbf{W}\mathbf{W}^T)^{-1}\mathbf{u}\Big\}\Bigg)\Bigg).$$

Claim: $f^*(D^*_{\mathbf{W},\mathbf{u}}(\mathbf{x}))$ is a strictly convex function of $\mathbf{x}$.

To show this claim, note that the following expression is a strongly-convex quadratic function of $\mathbf{x}$, since we have assumed that the spectral norm of $\mathbf{W}$ is bounded as $\sigma_{\max}(\mathbf{W}) \leq 1 < \sigma$:

$$\Big\{\frac{1}{2}\mathbf{x}^T((\mathbf{W}^T\mathbf{W})^{-1} - \sigma^{-2}I)\mathbf{x} - \mathbf{u}^T(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{x} + \mathbf{u}^T(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{u}\Big\}.$$

For simplicity, we denote the above strongly-convex function with $g(\mathbf{x})$ and define the function $h : \mathbb{R} \to \mathbb{R}$ as

$$h(y) := f^*\big(f'\big(\sqrt{\frac{\det(\mathbf{W}\mathbf{W}^T)}{\sigma^{2k}}} \times e^y\big)\big).$$

According to the above definitions, $f^*(D^*_{\mathbf{W},\mathbf{u}}(\mathbf{x})) = h(g(\mathbf{x}))$ is the composition of h and strongly-convex g. Note that h is a monotonically increasing function, since defining $c = \sqrt{\frac{\det(\mathbf{W}\mathbf{W}^T)}{\sigma^{2k}}} > 0$ we have

$$h'(y) = (c\,e^y)^2\,f''(c\,e^y) \geq 0, \tag{16}$$

which follows from the equality

$$f^*(f'(z)) := \sup_u\{uf'(z) - f(u)\} \;=\; zf'(z) - f(z)$$

that is a consequence of the definition of Fenchel-conjugate, implying that $\frac{\mathrm{d}f^*(f'(z))}{\mathrm{d}z} = zf''(z)$ for the convex f. Note that $h'(y) > 0$ holds everywhere, because f is assumed to be strictly convex. This proves that h is strictly increasing. Furthermore, h is a convex function, because $h'(y)$ is non-decreasing due to the assumption that $t^2f''(t)$ is non-decreasing over $t \in (0, +\infty)$. As a result, h is an increasing convex function.

Therefore, $f^*(D^*_{\mathbf{W},\mathbf{u}}(\mathbf{x})) = h(g(\mathbf{x}))$ is a composition of a strongly-convex g and an increasing convex $h : \mathbb{R} \to \mathbb{R}$. Therefore, as a well-known result in convex optimization (Boyd & Vandenberghe, 2004), the claim is true and $f^*(D^*_{\mathbf{W},\mathbf{u}}(\mathbf{x}))$ is a strictly convex function of $\mathbf{x}$.

We showed that the claim is true for every feasible $\mathbf{W}, \mathbf{u}$. Now, we prove that the pair $(G_{\mathbf{W},\mathbf{u}}, D^*_{\mathbf{W},\mathbf{u}})$ will not be a local Nash equilibrium for any feasible $\mathbf{W}, \mathbf{u}$. If the pair $(G_{\mathbf{W},\mathbf{u}}, D^*_{\mathbf{W},\mathbf{u}})$ was a local Nash equilibrium, $\mathbf{W}, \mathbf{u}$ would be a local minimum for the following minimax objective

Figure 18. SN-GAN's generated samples the first and last iterations of the main text's Figure 5.

Figure 19. WGAN-WC's generated samples at the first and last iterations of Figure 13.

where D^* is fixed to be $D^*_{\mathbf{W},\mathbf{u}}$:

$$\mathbb{E}\big[D^*(\mathbf{X})\big] - \mathbb{E}\big[f^*\big(D^*(\mathbf{WZ}+\mathbf{u})\big)\big]. \tag{17}$$

However, as shown earlier, for any feasible $\mathbf{W}, \mathbf{u}$, $f^*\big(D^*(\mathbf{x})\big)$ is a strictly-convex function of $\mathbf{x}$, which in turn shows that (17) is a strictly-concave function of variables $\mathbf{W}, \mathbf{u}$. This result shows that the objective can only take a local minimum on the boundary of the constrained set of $\|\mathbf{u}\|_2 \leq t$ and $\sigma_{\max}(\mathbf{W}) \leq 1$, where the local minimum cannot be a convex combination of any two feasible points. Therefore, $\|\mathbf{u}^*\|_2 = t$ and all singular values of $\mathbf{W}$ should be equal to 1 implying that $\mathbf{WW}^T = I$. However, note that $\nabla_{\mathbf{u}}\mathbb{E}[f^*(D^*(\mathbf{WZ}+\mathbf{u}))] = \mathbb{E}[h'(g(\mathbf{WZ}+\mathbf{u}))\nabla_{\mathbf{u}}g(\mathbf{WZ}+\mathbf{u})]$. This implies that

$$\begin{aligned}
&\mathbf{u}^T\nabla_{\mathbf{u}}\mathbb{E}[f^*(D^*(\mathbf{WZ}+\mathbf{u}))]\\
=&\,\mathbf{u}^T\mathbb{E}[h'(g(\mathbf{WZ}+\mathbf{u}))\nabla_{\mathbf{u}}g(\mathbf{WZ}+\mathbf{u})]\\
=&\,\mathbf{u}^T\mathbb{E}[h'(-\frac{1}{2\sigma^2}\|\mathbf{WZ}+\mathbf{u}\|_2^2+\frac{1}{2}\|\mathbf{Z}\|_2^2)\\
&\times((\mathbf{W}^T\mathbf{W})^{-1}-\sigma^{-2}I)\mathbf{WZ}]\\
&-\sigma^{-2}\|\mathbf{u}\|_2^2\mathbb{E}[h'(-\frac{1}{2\sigma^2}\|\mathbf{WZ}+\mathbf{u}\|_2^2+\frac{1}{2}\|\mathbf{Z}\|_2^2)]\\
<&\,0.
\end{aligned}$$

In the above equation, the last equality holds since we have

$$g(\mathbf{WZ}+\mathbf{u}) = -\frac{1}{2\sigma^2}\|\mathbf{WZ}+\mathbf{u}\|_2^2+\frac{1}{2}\|\mathbf{Z}\|_2^2.$$

Also, the last inequality holds because we showed h is a strictly increasing function and therefore

$$\mathbb{E}[h'(-\frac{1}{2\sigma^2}\|\mathbf{WZ}+\mathbf{u}\|_2^2+\frac{1}{2}\|\mathbf{Z}\|_2^2)] > 0.$$

Furthermore,

$$\begin{aligned}
&-\frac{1}{2\sigma^2}\|\mathbf{WZ}+\mathbf{u}\|_2^2+\frac{1}{2}\|\mathbf{Z}\|_2^2\\
=&\,\frac{1}{2}\mathbf{Z}^T(I-\sigma^{-2}(\mathbf{WW}^T))\mathbf{Z}-\frac{1}{\sigma^2}\mathbf{u}^T\mathbf{WZ}\\
=&\,\frac{1-\sigma^{-2}}{2}\|\mathbf{Z}\|_2^2-\frac{1}{\sigma^2}\mathbf{u}^T\mathbf{WZ}
\end{aligned}$$

Here $I-\sigma^{-2}(\mathbf{WW}^T)$ is a positive definite matrix which on the boundary local solution would be equal to $(1-\sigma^{-2})I$. Note that the jointly Gaussian variables $\mathbf{u}^T\mathbf{WZ}$ and $\mathbf{u}^T((\mathbf{W}^T\mathbf{W})^{-1}-\sigma^{-2}I)\mathbf{WZ} = (1-\sigma^{-2})\mathbf{u}^T\mathbf{WZ}$ are zero-mean and positively-correlated with correlation coefficient $+1$. Therefore, defining $Z' := (1-\sigma^{-2})\mathbf{u}^T\mathbf{WZ}$ we have

$$\begin{aligned}
&\mathbf{u}^T\mathbb{E}[h'(-\frac{1}{2\sigma^2}\|\mathbf{WZ}+\mathbf{u}\|_2^2+\frac{1}{2}\|\mathbf{Z}\|_2^2)\\
&\times((\mathbf{W}^T\mathbf{W})^{-1}-\sigma^{-2}I)\mathbf{WZ}]\\
=&\,\mathbb{E}[Z'h'(\alpha_0 Z'^2-\alpha_1 Z'+g_0(\tilde{\mathbf{Z}}))]
\end{aligned}$$

where $\alpha_0, \alpha_1 > 0$ and $g_0(\mathbf{Z})$ is a quadratic function of the $\tilde{\mathbf{Z}}$ components which are orthogonal to and independent from Z'. However, for any observation of $\tilde{\mathbf{Z}} = \tilde{\mathbf{z}}$ we have

$$\mathbb{E}[Z'h'(\alpha_0 Z'^2-\alpha_1 Z'+g_0(\tilde{\mathbf{Z}}))|\tilde{\mathbf{Z}}=\tilde{\mathbf{z}}] \leq 0.$$

Figure 20. WGAN-GP's generated samples at the first and last iterations of Figure 14.

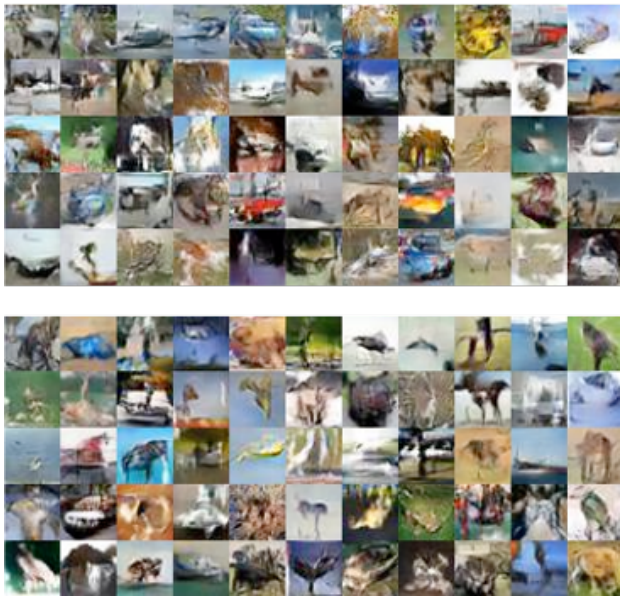

Figure 21. The images generated by the SN-GAN (DIM=128) trained on CIFAR-10 data with (top) ordinary and (bottom) proximal training.

The above holds because $h'(\alpha_0 Z'^2 - \alpha_1 Z' + g_0(\tilde{\mathbf{z}}))$ is a non-negative function of Z' that is symmetric around the positive $\frac{\alpha_1}{2\alpha_0}$ and is decreasing for $Z' < \frac{\alpha_1}{2\alpha_0}$ and increasing for $Z' > \frac{\alpha_1}{2\alpha_0}$. As a result, we have

$$\mathbf{u}^T \nabla_{\mathbf{u}} \mathbb{E}[f^*(D^*(\mathbf{WZ}+\mathbf{u}))] < 0$$

which implies that $-\mathbb{E}[f^*(D^*(\mathbf{WZ}+\mathbf{u}))]$ and hence the minimax objective will further decrease as we move from $\mathbf{u}^*$ toward $\mathbf{0}$. This result contradicts $\mathbf{u}$ being a local minimum of the minimax objective over $\|\mathbf{u}\|_2 \leq t$. Due to the shown contradiction, a pair with the form $(G_{\mathbf{W},\mathbf{u}}, D^*_{\mathbf{W},\mathbf{u}})$ cannot be a local Nash equilibrium in parameters $\mathbf{W}, \mathbf{u}$. Consequently, the minimax problem has no pure Nash equilibrium solutions, since in a pure Nash equilibrium the discriminator will be by definition optimal against the choice of generator.

Proof for W2GANs:

Consider the W2GAN problem with the assumed generator function:

$$\min_{\mathbf{W},\mathbf{u}:\, \|\mathbf{W}\|_2 \leq 1} \max_{D \text{ c-concave}} \mathbb{E}\big[D(\mathbf{X})\big] - \mathbb{E}\big[D^c(\mathbf{WZ}+\mathbf{u})\big], \tag{18}$$

where the c-transform is defined for the quadratic cost function $c(\mathbf{x},\mathbf{x}') = \frac{1}{2}\|\mathbf{x}-\mathbf{x}'\|_2^2$. Similar to the f-GAN case, define $D^*_{\mathbf{W},\mathbf{u}}$ to be the optimal discriminator for the generator function parameterized by $\mathbf{W}, \mathbf{u}$. Note that $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$ and $\mathbf{WZ}+\mathbf{u} \sim \mathcal{N}(\mathbf{u}, \mathbf{WW}^T)$.

According to the Brenier's theorem (Villani, 2008), the optimal transport from the Gaussian data distribution to the Gaussian generative model will be

$$\psi^{\text{opt}}(\mathbf{x}) = \mathbf{x} - \nabla_{\mathbf{x}} D^*_{\mathbf{W},\mathbf{u}}(\mathbf{x}).^{1}$$

As a well-known result regarding the second-order optimal transport map between two Gaussian distributions, the optimal transport will be a linear transformation as $\psi^{\text{opt}}(\mathbf{x}) = \frac{1}{\sigma}(\mathbf{WW}^T)^{1/2}\mathbf{x} + \mathbf{u}$. This result shows that

$$\nabla_{\mathbf{x}} D^*_{\mathbf{W},\mathbf{u}}(\mathbf{x}) = \big(I - \frac{1}{\sigma}(\mathbf{WW}^T)^{1/2}\big)\mathbf{x} - \mathbf{u}. \tag{19}$$

Note that the c-transform for cost $c(\mathbf{x},\mathbf{x}') = \frac{1}{2}\|\mathbf{x}-\mathbf{x}'\|_2^2$ satisfies $D^c(\mathbf{x}) = (\frac{1}{2}\|\mathbf{x}\|^2 - D(\mathbf{x}))^* - \frac{1}{2}\|\mathbf{x}\|^2$ where g^* denotes g's Fenchel-conjugate. For general convex quadratic function $g(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} + \mathbf{b}^T\mathbf{x}$ we have $g^*(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\mathbf{b})^T A^\dagger(\mathbf{x}-\mathbf{b})$ where $A^\dagger$ denotes A's Moore Penrose pseudoinverse. Therefore, for the c-transform of the optimal discriminator we will have

$$\nabla_{\mathbf{x}} D^{*^c}_{\mathbf{W},\mathbf{u}}(\mathbf{x}) = \big(\sigma((\mathbf{WW}^T)^{1/2})^\dagger - I\big)\mathbf{x} - \sigma((\mathbf{WW}^T)^{1/2})^\dagger \mathbf{u}.$$

Since every feasible $\mathbf{W}$ satisfies the bounded spectral norm condition as $\sigma_{\max}(\mathbf{W}) \leq 1 < \sigma$, the optimal $D^{*^c}_{\mathbf{W},\mathbf{u}}$ will be a quadratic function whose Hessian has at least one strictly positive eigenvalue along the principal eigenvector of $\mathbf{WW}^T$. The positive eigenvalue exists in general

[1] Notice the change of variable $D(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2 - \psi(\mathbf{x})$ compared to the formulation discussed at (Villani, 2008; Feizi et al., 2017) which are based on the function ψ.

Figure 22. The images generated by the WGAN-WC (DIM=128) trained on CIFAR-10 data with (top) ordinary and (bottom) proximal training.

Figure 23. The images generated by the WGAN-WC (DIM=64) trained on CIFAR-10 data with (top) ordinary and (bottom) proximal training.

case where $\mathbf{Z}$'s dimension can be even smaller than $\mathbf{X}$'s dimension. If we had the stronger assumption that the two dimensions exactly match, similar to the f-GAN problem considered, then the pseudo-inverse $A^{\dagger}$ would be the same as the inverse A^{-1} resulting in a strongly-convex quadratic $D^{*c}_{\mathbf{W},\mathbf{u}}$. Nevertheless, as we prove here, the theorem's result on W2GAN holds in the general case and does not necessarily require the same dimension between $\mathbf{X}$ and $\mathbf{Z}$.

Consider the W2GAN minimax objective for the pair $(G_{\mathbf{W},\mathbf{u}}, D^*)$ where D^* is fixed to be the optimal $D^*_{\mathbf{W},\mathbf{u}}$:

$$\mathbb{E}\big[D(\mathbf{X})\big] - \mathbb{E}\big[D^{*c}(\mathbf{WZ}+\mathbf{u})\big], \tag{20}$$

If $(G_{\mathbf{W},\mathbf{u}}, D^*_{\mathbf{W},\mathbf{u}})$ was a local Nash equilibrium, the variables $\mathbf{W}, \mathbf{u}$ would provide a local minimum to the above objective. However, since $D^{*c}_{\mathbf{W},\mathbf{u}}$ is shown to be a quadratic function with a Hessian possessing positive eigenvalues, the above minimax objective will not have a local minimum inside the feasible set $\|\mathbf{u}\|_2 < t$. Also, note that

$$\begin{aligned}
&\mathbf{u}^T \nabla_{\mathbf{u}} \mathbb{E}[D^{*c}(\mathbf{WZ}+\mathbf{u})] \\
&= -\|\mathbf{u}\|_2^2 + \sigma \mathbb{E}\big[\mathbf{u}^T((\mathbf{W}\mathbf{W}^T)^{1/2})^{\dagger} - I)\mathbf{WZ}\big] \\
&= -\|\mathbf{u}\|_2^2 \\
&< 0.
\end{aligned}$$

As a result, the minimax objective cannot have a local minimum over the boundary $\|\mathbf{u}\|_2 = t$ because the objective will be strictly decreasing when we move toward the origin. Due to the shown contradiction, the minimax problem possesses no local Nash equilibrium solutions with the form $(G_{\mathbf{W},\mathbf{u}}, D^*_{\mathbf{W},\mathbf{u}})$ and therefore no pure Nash equilibrium solutions.

For the parameterized case with a quadratic discriminator $D_{A,\mathbf{b}}(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} + \mathbf{b}^T\mathbf{x}$, first of all note that as shown in the proof the optimal discriminator $D^*_{\mathbf{W},\mathbf{u}}$ for any generator parameter $\mathbf{W}, \mathbf{u}$ will be a c-concave quadratic function. Therefore, the optimal solution for the discriminator does not change because of the new quadratic constraint. Furthermore, the discriminator optimization problem has a concave objective in parameters $A, \mathbf{b}$. This is because the discriminator $D_{A,\mathbf{b}}(\mathbf{x})$ is a linear function in terms of $A, \mathbf{b}$, and $D^c_{A,\mathbf{b}}(\mathbf{x}) = \sup_{\mathbf{x}'} D_{A,\mathbf{b}}(\mathbf{x}') - c(\mathbf{x}', \mathbf{x})$ is a convex function of $A, \mathbf{b}$ as the supremum of some affine functions is convex.

As a result, the discriminator optimization reduces to maximizing a concave objective of $A, \mathbf{b}$ constrained to a convex set $\{A : I - A \succcurlyeq 0\}$ which is equivalent to the c-concave constraint on the quadratic $D_{A,\mathbf{b}}$. Hence, any local solution to this optimization problem will also be a global solution. This result implies that any local Nash equilibrium for the new parameterized minimax problem will have the form $(G_{\mathbf{W},\mathbf{u}}, D^*_{\mathbf{W},\mathbf{u}})$, which as we have already shown does not exist under the theorem's assumptions.

Proof for the 1-dimensional WGAN:

Consider the 1-dimensional Wasserstein GAN problem for the assumed linear generator function:

$$\min_{w,u:\,|w|\le 1} \max_{\|D\|_{\mathrm{Lip}}\le 1} \mathbb{E}[D(X)] - \mathbb{E}[D(wZ+u)]. \tag{21}$$

Figure 24. The images generated by the SN-GAN trained on CelebA data with (top) ordinary and (bottom) proximal training.

Figure 25. The images generated by the WGAN-WC trained on CelebA data with (top) ordinary and (bottom) proximal training.

The inner maximization problem can be rewritten as

$$\max_{\|D\|_{\text{Lip}}\leq 1} \int \big(p_X(x) - p_{wZ+u}(x)\big)\, D(x)\,\mathrm{d}x. \qquad (22)$$

Here we have

$$\begin{aligned} &p_X(x) - p_{wZ+u}(x) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{\frac{-1}{2\sigma^2}x^2\Big\} - \frac{1}{\sqrt{2\pi w^2}} \exp\Big\{\frac{-1}{2w^2}(x-u)^2\Big\} \end{aligned}$$

Since $|w| \leq 1 < \sigma$, it can be seen that the above difference will be positive everywhere except over an interval (a_1, a_2), where a_1, a_2 are the two solutions to the quadratic equation:

$$\big(\frac{1}{w^2} - \frac{1}{\sigma^2}\big)x^2 - 2\frac{u}{w^2}x + \big(\frac{u^2}{w^2} - \log(\frac{\sigma}{|w|})\big) = 0. \quad (23)$$

Note that the above quadratic equation has two distinct solutions $a_1 < a_2$, since $|w| < \sigma$ and $\log(\frac{\sigma}{|w|}) > 0$ leading to the positive discriminant:

$$4\frac{u^2}{w^2\sigma^2} + 4\log(\frac{\sigma}{|w|})\big(\frac{1}{w^2} - \frac{1}{\sigma^2}\big) \;>\; 0. \qquad (24)$$

As the function D in the maximization problem (22) is only constrained to be 1-Lipschitz, the optimal $D^*_{w,u}$'s slope must be equal to -1 over $(-\infty, a_1]$ and equal to $+1$ over $[a_2, \infty)$, in order to allow the maximum increase in the maximization objective. Over the interval (a_1, a_2), we claim that for the optimal D is a convex function, because otherwise its double Fenchel-conjugate D^{**}, which is by definition convex, achieves a higher value.

First of all, note that the double Fenchel-conjugate D^{**} will not be different from D outside the (a_1, a_2) interval, because D^{**} is defined to provide the largest convex function satisfying $D^{**} \leq D$, and D is supposed to be 1-Lipschitz taking its minimum derivative on $(-\infty, a_1]$ and its maximum derivative over $[a_2, \infty)$. Next, since D^{**} lower-bounds D, it results in a non-smaller integral value over the interval (a_1, a_2) as $p_X(x) - p_{wZ+u}(x)$ takes negative values over (a_1, a_2). If D is not convex, then D^{**} provides a strict lower-bound for D which matches D over $(\infty, a_1] \cup [a_2, \infty)$. Therefore, the convex 1-Lipschitz D^{**} results in a greater objective that is a contradiction to D's optimality. This contradiction proves that the optimal discriminator $D^*_{w,u}$ is a convex function.

Also, note that since $D^*_{w,u}$ is constrained to be 1-Lipschitz where it takes its minimum and maximum derivative values over $(-\infty, a_1]$ and $[a_2, \infty)$ it will be lower-bounded as

$$D^*_{w,u}(x) \geq \max\{D^*_{w,u}(a_1)-(x-a_1), D^*_{w,u}(a_2)+(x-a_2)\}.$$

Since the above lower-bound matches $D^*_{w,u}(x)$ over $(-\infty, a_1] \cup [a_2, \infty)$ and $D^*_{w,u}$ is supposed to maximize the inner product with $p_X(x) - p_{wZ+u}(x)$ which is negative over (a_1, a_2), the above lower-bound is in fact tight and equal to $D^*_{w,u}$. Hence, not only the optimal $D^*_{w,u}$ is convex, but also for two real constants a, b we have

$$D^*_{w,u}(x) = |x - a| + b.$$

Since the minimax objective does not change by adding a constant to the discriminator function, for every feasible $|w| \leq 1$ there exists an optimal solution $D^*_{w,u}$ for (22) that is a convex function of the form $D^*_{w,u}(x) = |x - a|$. This result proves that if a local Nash equilibrium of the form $(G_{w,u}, D^*_{w,u})$ existed for the WGAN problem, then the optimal u would be on the boundary $|u| = t$. However, note that $D^*_{w,u}$ is a convex function with a non-constant increasing derivative. Also, note that for the optimal parameter a^* such that $D^*_{w,u}(x) = |x - a^*|$ we must have $|a^*| > |u|$. This is because the derivative of the following discriminator optimization with respect to a is

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}a} \int \big(p_X(x) - p_{wZ+u}(x)\big)\, |x - a|\, \mathrm{d}x \\
&= \int \big(-p_X(x) + p_{wZ+u}(x)\big)\, \operatorname{sign}(x - a)\, \mathrm{d}x \\
&= 2\Pr(wZ + u \geq a) - 2\Pr(X \geq a) \qquad (25)
\end{aligned}$$

which will be greater than $1 - 2\Pr(X \geq a) > 0$ if $0 < a < u$ or will be less than $2(\Pr(Z \geq \frac{a-u}{w}) - \Pr(X \geq a)) < 2(\frac{1}{2} - \frac{1}{2}) = 0$ if $u < a < 0$. Therefore, for the optimal a^* we have $|a^*| > |u|$ and fixing the optimal discriminator the second term of the minimax objective $-\mathbb{E}[|wZ + u - a^*|]$ will take smaller values if we move u closer to 0. The minimax objective will further decrease as we move u toward 0 which is in the feasible set and gives a contradiction. Therefore, the WGAN problem has no local Nash equilibria with the form $(G_{w,u}, D^*_{w,u})$, because if $(G_{w,u}, D^*_{w,u})$ was a local Nash equilibrium then w, u would be a local minimum for the following objective where D^* is fixed to be $D^*_{w,u}$ which as we showed cannot exist. This shows that the WGAN problem does not have a Nash equilibrium and completes the proof for the WGAN case. □

Remark 1. *Consider the same setting as in Theorem 1. However, unlike Theorem 1 suppose that $\sigma < 1$ and $\sigma_{\min}(\mathbf{W}) \geq 1$ where $\sigma_{\min}$ stands for the minimum singular value. Then, for the WGAN and W2GAN problems described in Theorem 1, the Wasserstein distance-minimizing generator results in a Nash equilibrium.*

Proof. **Proof for the W2GAN:**

For the W2GAN case, note that if we repeat the same steps as in the proof of Theorem 1, we can show

$$\begin{aligned}
\nabla_{\mathbf{x}} D^{*^c}_{\mathbf{W},\mathbf{u}}(\mathbf{x}) = &\big(\sigma(\mathbf{W}\mathbf{W}^T)^{-1/2} - I\big)\mathbf{x} \qquad (26)\\
&- \sigma(\mathbf{W}\mathbf{W}^T)^{-1/2}\mathbf{u}.
\end{aligned}$$

which is a concave quadratic function of $\mathbf{x}$, since the assumptions imply that $(\mathbf{W}\mathbf{W}^T)^{-1} \preccurlyeq \sigma^{-2} I$. Here $\mathbf{W}$ is supposed to be a full-rank square matrix as its minimum singular value is assumed to be positive and $\mathbf{Z}$ has the same dimension as $\mathbf{X}$.

We claim that for the feasible choice $\mathbf{W}^* = I$ and $\mathbf{u}^* = \mathbf{0}$, the pair $(G_{\mathbf{W}^*,\mathbf{u}^*}, D^*_{\mathbf{W}^*,\mathbf{u}^*})$ results in a Nash equilibrium of the minimax problem. Considering the definition of the optimal discriminator $D^*_{\mathbf{W}^*,\mathbf{u}^*}$, its optimlaity for $G_{\mathbf{W}^*,\mathbf{u}^*}$ directly follows. Moreover, (26) implies that

$$D^{*^c}_{\mathbf{W}^*,\mathbf{u}^*}(\mathbf{x}) = \frac{\sigma - 1}{2}\|\mathbf{x}\|_2^2. \qquad (27)$$

As a result, fixing the above discriminator function the minimax objective will be

$$\begin{aligned}
&\mathbb{E}[D^*(\mathbf{X})] - \mathbb{E}\big[\frac{\sigma - 1}{2}\|\mathbf{W}\mathbf{Z} + \mathbf{u}\|_2^2\big] \\
&= \mathbb{E}[D^*(\mathbf{X})] + \frac{1 - \sigma}{2}\left(\|\mathbf{W}\|_F^2 + \|\mathbf{u}\|_2^2\right)
\end{aligned}$$

which is minimized at $\mathbf{W} = I$ and $\mathbf{u} = \mathbf{0}$ over the specified feasible set, as we know the Frobenius norm-squared, $\|\mathbf{W}\|_F^2$, is the sum of the squared of $\mathbf{W}$'s singular values. Therefore, the claim holds and the choice $\mathbf{W}^* = I$ and $\mathbf{u}^* = 0$ results in the optimal solution and a Nash equilibrium.

Proof for the 1-dimensional Wasserstein GAN:

Here we select the parameters $w^* = 1$, $u^* = 0$. We claim that the optimal discriminator function for this choice is the negative absolute value function $D^*(x) = -|x|$. Note that the optimal 1-Lipschitz D^* solves the following problem:

$$\max_{\|D\|_{\mathrm{Lip}} \leq 1} \int \big(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} - \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\big) D(x)\, \mathrm{d}x. \quad (28)$$

In the above objective given $\eta = \sqrt{\frac{2\sigma^2 \log(1/\sigma)}{1-\sigma^2}}$, the function $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} - \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ is positive over $(-\eta, \eta)$ and negative elsewhere. Therefore, the optimal D^* should get the maximum $+1$ derivative over $(-\infty, -\eta]$ and the minimum -1 derivative over $[+\eta, +\infty)$. Because of the even structure of $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} - \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, there exists an even optimal D^* because $\frac{D^*(x)+D^*(-x)}{2}$ remains 1-Lipschitz and optimal for any optimal 1-Lipschitz discriminator D^*.

The optimal even D^* should further be continuous as a 1-Lipschitz function, implying that such a D^* is decreasing over $(0, \eta]$ and increasing over $[-\eta, 0)$. Enforcing the maximum derivative over the two interval results in the optimal $D^*(x) = -|x|$.

Therefore, $D^*(x) = -|x|$ provides an optimal discriminator for $w^* = 1$, $u^* = 0$. Also, for this D^* the minimax objective of the Wasserstein GAN will be

$$\mathbb{E}[-|X|] - \mathbb{E}[-|wZ + u|] \ = -\mathbb{E}[|X|] + \mathbb{E}[|wZ + u|]$$

In the above equation, $wZ + u \sim \mathcal{N}(u, w^2)$, showing that the above objective is minimized at $w^* = 1, u = 0$ considering the assumed feasible set where $|w| \geq 1$. As a result, the pair (G_{w^*,u^*}, D^*) provides a Nash equilibrium to the WGAN minimax game. □

3.3. Proof of Proposition 2

Proposition 2. *(G^*, D^*) is a λ-proximal equilibrium if and only if for every $G \in \mathcal{G}$, $D \in \mathcal{D}$ we have*

$$\begin{aligned} V(G^*, D) \ &\leq \ V(G^*, D^*) \qquad (29)\\ &\leq \max_{\widetilde{D} \in \mathcal{D}} V(G, \widetilde{D}) - \lambda \big\| \widetilde{D} - D^* \big\|^2. \end{aligned}$$

Proof. **Proof of the $\Rightarrow$ direction:**

Assume that (G^*, D^*) is a λ-proximal equilibrium. According to the definition of the proximal equilibrium, the following holds for every $G \in \mathcal{G}$ and $D \in \mathcal{D}$:

$$V_\lambda^{\text{prox}}(G^*, D) \ \leq \ V_\lambda^{\text{prox}}(G^*, D^*) \ \leq \ V_\lambda^{\text{prox}}(G, D^*). \qquad (30)$$

Claim: $V_\lambda^{\text{prox}}(G^*, D^*) = V(G^*, D^*)$.

To show this claim, note that

$$V_\lambda^{\text{prox}}(G^*, D^*) := \max_{\widetilde{D} \in \mathcal{D}} V(G^*, \widetilde{D}) - \lambda \big\| \widetilde{D} - D^* \big\|^2. \qquad (31)$$

In this optimization, the optimal solution $\tilde{D}$ is D^* itself. Otherwise, for the optimal $\tilde{D} \in \mathcal{D}$ we have $\|\tilde{D} - D^*\| > 0$ and as a result

$$V_\lambda^{\text{prox}}(G^*, D^*) < V(G^*, \tilde{D}) \leq V_\lambda^{\text{prox}}(G^*, \tilde{D}), \qquad (32)$$

which is a contradiction given that (G^*, D^*) is a λ-proximal equilibrium. Therefore, D^* optimizes the proximal optimization, which shows the claim is valid and we have $V_\lambda^{\text{prox}}(G^*, D^*) = VG^*, D^*)$. Knowing that $V(G, D) \leq V_\lambda^{\text{prox}}(G, D)$ holds for every $G \in \mathcal{G}, D \in \mathcal{D}$, we have

$$\begin{aligned} V(G^*, D) &\leq V_\lambda^{\text{prox}}(G^*, D)\\ &\leq V_\lambda^{\text{prox}}(G^*, D^*)\\ &= V(G^*, D^*). \end{aligned}$$

Furthermore,

$$\begin{aligned} V(G^*, D^*) &= V_\lambda^{\text{prox}}(G^*, D^*) \qquad (33)\\ &\leq V_\lambda^{\text{prox}}(G, D^*) \qquad (34)\\ &= \max_{\widetilde{D} \in \mathcal{D}} V(G, \widetilde{D}) - \lambda \big\| \widetilde{D} - D^* \big\|^2. \end{aligned}$$

Therefore, the proof is complete.

Proof of the $\Leftarrow$ direction:

Suppose that for (G^*, D^*) the following holds for every $G \in \mathcal{G}$ and $D \in \mathcal{D}$:

$$\begin{aligned} V(G^*, D) \ &\leq \ V(G^*, D^*) \qquad (35)\\ &\leq \max_{\widetilde{D} \in \mathcal{D}} V(G, \widetilde{D}) - \lambda \big\| \widetilde{D} - D^* \big\|^2. \end{aligned}$$

We claim that $V(G^*, D^*) = V_\lambda^{\text{prox}}(G^*, D^*)$. To show this claim, consider the definition of the λ-proximal equilibrium:

$$V_\lambda^{\text{prox}}(G^*, D^*) := \max_{\tilde{D} \in \mathcal{D}} V(G^*, \tilde{D}) - \lambda \big\| D^* - \tilde{D} \big\|. \qquad (36)$$

Here D^* maximizes the objective because we have assumed that $V(G^*, D) \ \leq \ V(G^*, D^*)$ holds for every $D \in \mathcal{D}$. Therefore, the claim is valid and $V(G^*, D^*) = V_\lambda^{\text{prox}}(G^*, D^*)$.

Also, note that for every D the solution $\tilde{D}$ in the proximal optimization satisfies $V_\lambda^{\text{prox}}(G^*, D) \leq V(G^*, \tilde{D})$. Combining these results with (35), we obtain the following inequalities which hold for every $G \in \mathcal{G}$ and $D \in \mathcal{D}$:

$$V_\lambda^{\text{prox}}(G^*, D) \ \leq \ V_\lambda^{\text{prox}}(G^*, D^*) \ \leq \ V_\lambda^{\text{prox}}(G, D^*). \qquad (37)$$

The above equation shows that the pair (G^*, D^*) is a λ-proximal equilibrium. □

3.4. Proof of Proposition 3

Proposition 3. *Define $\text{PE}_\lambda(V)$ to be the set containing the λ-proximal equilibria of $V(G, D)$. Then, if $\lambda_1 \leq \lambda_2$ we have*

$$\text{PE}_{\lambda_2}(V) \subseteq \text{PE}_{\lambda_1}(V). \qquad (38)$$

Proof. Consider a λ_2-proximal equilibrium (G^*, D^*). As a result of Proposition 2, for every $G \in \mathcal{G}$ and $D \in \mathcal{D}$ we have

$$\begin{aligned} V(G^*, D) \ &\leq \ V(G^*, D^*) \qquad (39)\\ &\leq \max_{\widetilde{D} \in \mathcal{D}} V(G, \widetilde{D}) - \lambda_2 \big\| \widetilde{D} - D^* \big\|^2. \end{aligned}$$

Since $\lambda_1 \leq \lambda_2$, the following holds

$$\max_{\widetilde{D} \in \mathcal{D}} V(G, \widetilde{D}) - \lambda_2 \big\| \widetilde{D} - D^* \big\|^2 \qquad (40)$$

$$\leq \max_{\widetilde{D}\in\mathcal{D}} V(G,\widetilde{D}) - \lambda_1\big\|\widetilde{D}-D^*\big\|^2,$$

which shows that

$$\begin{aligned} V(G^*,D) &\leq V(G^*,D^*) \\ &\leq \max_{\widetilde{D}} V(G,\widetilde{D}) - \lambda_1\big\|\widetilde{D}-D^*\big\|^2. \end{aligned} \tag{41}$$

Due to Proposition 2, (G^*, D^*) will be a λ_1-proximal equilibrium as well. Hence, the proof is complete and we have

$$\mathrm{PE}_{\lambda_2}(V) \subseteq \mathrm{PE}_{\lambda_1}(V). \tag{42}$$

□

3.5. Proof of Proposition 4

Proposition 4. *Consider the maximization problem in the definition of proximal operator where generator $G_{\boldsymbol{\theta}}$ and discriminator $D_{\mathbf{w}}$ have been parameterized by vectors $\boldsymbol{\theta}$, $\mathbf{w}$, respectively. Suppose that*

- *For the considered discriminator norm $\|\cdot\|$, $\|D_{\mathbf{w}} - D\|^2$ is η_1-strongly convex in $\mathbf{w}$ for any function D, i.e. for any $\mathbf{w}, \mathbf{w}', D$:*

$$\Big\|\nabla_{\mathbf{w}}\|D_{\mathbf{w}}-D\|^2 - \nabla_{\mathbf{w}}\|D_{\mathbf{w}'}-D\|^2\Big\|_2 \geq \eta_1\big\|\mathbf{w}-\mathbf{w}'\big\|_2,$$

- *For every $G_{\boldsymbol{\theta}}$, The GAN minimax objective $V(G_{\boldsymbol{\theta}}, D_{\mathbf{w}})$ is η_2-smooth in $\mathbf{w}$, i.e.*

$$\big\|\nabla_{\mathbf{w}}V(G_{\boldsymbol{\theta}},D_{\mathbf{w}}) - \nabla_{\mathbf{w}}V(G_{\boldsymbol{\theta}},D_{\mathbf{w}})\big\|_2 \leq \eta_2\|\mathbf{w}-\mathbf{w}'\|_2.$$

Under the above assumptions, if $\eta_2 < \lambda\eta_1$, the proximal objective is $(\lambda\eta_1 - \eta_2)$-strongly concave. Then, the maximization problem has a unique solution $\mathbf{w}^*$ and if $V(G_{\boldsymbol{\theta}}, D_{\mathbf{w}})$ is differentiable with respect to θ then

$$\nabla_\theta V_\lambda^{\mathrm{prox}}(G_{\boldsymbol{\theta}}, D_{\mathbf{w}}) = \nabla_\theta V(G_{\boldsymbol{\theta}}, D_{\mathbf{w}^*}). \tag{43}$$

Proof. Consider the definition of a λ-proximal equilibrium in the parameterized space:

$$V_\lambda^{\mathrm{prox}}(G_{\boldsymbol{\theta}}, D_{\mathbf{w}}) := \max_{\widetilde{\mathbf{w}}} V(G_{\boldsymbol{\theta}}, D_{\widetilde{\mathbf{w}}}) - \lambda\big\|D_{\widetilde{\mathbf{w}}} - D_{\mathbf{w}}\big\|^2. \tag{44}$$

In the above optimization problem, the first term $V(G_{\boldsymbol{\theta}}, D_{\widetilde{\mathbf{w}}})$ is assumed to be η_2-smooth in $\widetilde{\mathbf{w}}$, while the second term $\lambda\|D_{\widetilde{\mathbf{w}}} - D_{\mathbf{w}}\|^2$ will be $\lambda\eta_1$-strongly convex in $\widetilde{\mathbf{w}}$. As a result, the sum of the two terms will be $(\lambda\eta_1 - \eta_2)$-strongly concave if $\eta_2 < \lambda\eta_1$ holds. Since the objective function is strongly-concave in $\widetilde{\mathbf{w}}$, it will be maximized by a unique solution $\mathbf{w}^*$. Moreover, applying the generalized Danskin's theorem (Bernhard & Rapaport, 1995) implies that the following holds at the optimal $\mathbf{w}^*$:

$$\nabla_\theta V_\lambda^{\mathrm{prox}}(G_{\boldsymbol{\theta}}, D_{\mathbf{w}}) = \nabla_\theta V(G_{\boldsymbol{\theta}}, D_{\mathbf{w}^*}). \tag{45}$$

□

3.6. Proof of Theorem 2

Theorem 2. *Consider the second-order Wasserstein GAN problem with a quadratic cost $c(\mathbf{x}, \mathbf{x}') = \eta\|\mathbf{x} - \mathbf{x}'\|_2^2$. Suppose that the set of optimal discriminators $\{D^{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ is convex. Then, $(G_{\boldsymbol{\theta}^*}, D^{\boldsymbol{\theta}^*})$ for the Wasserstein distance-minimizing generator $G_{\boldsymbol{\theta}^*} \in \mathcal{G}$ will provide a $\frac{1}{4\eta}$-proximal equilibrium with respect to the Sobolev norm $\|\cdot\|_{\dot{H}^1}$.*

Proof.

Lemma 2. *Suppose that f is a γ-strongly convex function according to norm $\|\cdot\|$, i.e. for any $x, y \in \mathrm{dom}(f)$ and $\lambda \in [0,1]$ we have*

$$\begin{aligned} f(\lambda x + (1-\lambda)y) \leq& \lambda f(x) + (1-\lambda)f(y) \\ &- \frac{\gamma\lambda(1-\lambda)}{2}\|x-y\|^2. \end{aligned} \tag{46}$$

Consider the following optimization problem where the feasible set $\mathcal{X}$ is a convex set and x^* is the optimal solution,

$$\min_{x\in\mathcal{X}} f(x). \tag{47}$$

Then, for every $x \in \mathcal{X}$ we have

$$f(x) - f(x^*) \geq \frac{\gamma}{2}\|x - x^*\|^2. \tag{48}$$

Proof. If we apply the strong-convexity definition (46) to $x \in \mathcal{X}$, x^* we obtain

$$\begin{aligned} f(\lambda x + (1-\lambda)x^*) \leq& \lambda f(x) + (1-\lambda)f(x^*) \\ &- \frac{\gamma\lambda(1-\lambda)}{2}\|x-x^*\|^2. \end{aligned} \tag{49}$$

The above inequality results in

$$\begin{aligned} f(\lambda x + (1-\lambda)x^*) - f(x^*) \leq& \lambda(f(x) - f(x^*)) \\ &- \frac{\gamma\lambda(1-\lambda)}{2}\|x-x^*\|^2. \end{aligned} \tag{50}$$

Note that $\mathcal{X}$ is assumed to be a convex set and therefore $\lambda x + (1-\lambda)x^* \in \mathcal{X}$ implying $f(x^*) \leq f(\lambda x + (1-\lambda)x^*)$. As a result, for every $0 \leq \lambda \leq 1$

$$\frac{\gamma\lambda(1-\lambda)}{2}\|x-x^*\|^2 \leq \lambda(f(x) - f(x^*)). \tag{51}$$

Thus for every $0 < \lambda \leq 1$ we have

$$\frac{\gamma(1-\lambda)}{2}\|x-x^*\|^2 \leq f(x) - f(x^*), \tag{52}$$

which proves that $\frac{\gamma}{2}\|x-x^*\|^2 \leq f(x) - f(x^*)$ and completes Lemma 2's proof. □

Based on Proposition 2 and the definition of $D^{\boldsymbol{\theta}}$, we only need to show that for the W2GAN's objective, which we denote by $V(G, D)$, the following holds for every $\boldsymbol{\theta}$:

$$V(G_{\boldsymbol{\theta}^*}, D^{\boldsymbol{\theta}^*}) \leq \max_{D \in \mathcal{D}} V(G_{\boldsymbol{\theta}}, D) - \frac{1}{2\eta} \big\| D - D^{\boldsymbol{\theta}^*} \big\|_{\dot{H}^1}^2. \tag{53}$$

To show the above inequality, it suffices to prove the following inequality

$$V(G_{\boldsymbol{\theta}^*}, D^{\boldsymbol{\theta}^*}) \leq V(G_{\boldsymbol{\theta}}, D^{\boldsymbol{\theta}}) - \frac{1}{4\eta} \big\| D^{\boldsymbol{\theta}} - D^{\boldsymbol{\theta}^*} \big\|_{\dot{H}^1}^2. \tag{54}$$

Claim: For the W2GAN problem, we have

$$V(G_{\boldsymbol{\theta}}, D^{\boldsymbol{\theta}}) = \frac{1}{4\eta} \mathbb{E}\big[\|\nabla D^{\boldsymbol{\theta}}(\mathbf{X})\|_2^2\big].$$

To show this claim, note that according to the W2GAN's formulation we have $V(G_{\boldsymbol{\theta}}, D^{\boldsymbol{\theta}}) := W_c(P_{\mathbf{X}}, P_{G_{\boldsymbol{\theta}}(\mathbf{Z})})$ where $c(\mathbf{x}, \mathbf{x}') = \eta \|\mathbf{x} - \mathbf{x}'\|_2^2$ is the second-order cost function specified in the theorem. We start by proving this result for $\eta = \frac{1}{2}$. In this case, the Brenier's theorem (Ambrosio & Gigli, 2013) proves that the optimal transport map from the data variable $\mathbf{X}$ to the generative model $G_{\boldsymbol{\theta}}(\mathbf{Z})$ can be derived from the gradient of the optimal $D^{\boldsymbol{\theta}}$ as follows

$$\psi^{\text{opt}}(\mathbf{x}) = \mathbf{x} - \nabla D^{\boldsymbol{\theta}}(\mathbf{x}). \tag{55}$$

which plugged into the optimal transport objective $W_c(P, Q) := \inf_{\Pi(P,Q)} \mathbb{E}[c(\mathbf{X}, \mathbf{X}')]$ proves that

$$V(G_{\boldsymbol{\theta}}, D^{\boldsymbol{\theta}}) := W_c(P_{\mathbf{X}}, P_{G_{\boldsymbol{\theta}}(\mathbf{Z})}) = \mathbb{E}\big[\frac{1}{2}\|\nabla D^{\boldsymbol{\theta}}(\mathbf{X})\|_2^2\big].$$

The above equation proves the result holds for $\eta = \frac{1}{2}$. For a general $\eta > 0$, note that applying a simple change of variable in the Kantorovich duality representation and solving the dual problem for $\tilde{D}(\mathbf{x}) = 2\eta D(\mathbf{x})$ shows that $\psi^{\text{opt}}(\mathbf{x}) = \mathbf{x} - \frac{1}{2\eta}\nabla \tilde{D}^{\boldsymbol{\theta}}(\mathbf{x})$ transport samples from the data domain to the generative model. This is due to the fact that after applying this change of variable the Kantorovich duality reduces to η multiplied to the dual problem for $\eta = \frac{1}{2}$. As a result, applying the transport map to the definition of the optimal transport cost shows that

$$V(G_{\boldsymbol{\theta}}, D^{\boldsymbol{\theta}}) := W_c(P_{\mathbf{X}}, P_{G_{\boldsymbol{\theta}}(\mathbf{Z})}) = \frac{1}{4\eta} \mathbb{E}\big[\|\nabla D^{\boldsymbol{\theta}}(\mathbf{X})\|_2^2\big],$$

proving the claim holds for any $\eta > 0$.

Substituting the discriminator maximization with the result in the above claim, the W2GAN problem reduces to the following problem:

$$\min_{\boldsymbol{\theta}} \frac{1}{4\eta} \mathbb{E}\big[\big\|\nabla D^{\boldsymbol{\theta}}(\mathbf{X})\big\|_2^2\big]. \tag{56}$$

Here we can equivalently optimize for $D^{\boldsymbol{\theta}} \in \{D^{\boldsymbol{\theta}} : \theta \in \Theta\}$ instead of minimizing over the variable $\boldsymbol{\theta}$, obtaining

$$\min_{D^{\boldsymbol{\theta}}} \frac{1}{4\eta} \mathbb{E}\big[\big\|\nabla D^{\boldsymbol{\theta}}(\mathbf{X})\big\|_2^2\big]. \tag{57}$$

Note that the term $\frac{1}{2}\mathbb{E}\big[\big\|\nabla D^{\boldsymbol{\theta}}(\mathbf{X})\big\|_2^2\big] = \frac{1}{2}\big\|D^{\boldsymbol{\theta}}\big\|_{\dot{H}^1}^2$ reduces to the squared of the defined Sobolev norm in a semi-Hilbert space, which results in a 1-strongly convex function according to $\|\cdot\|_{\dot{H}^1}$ with strong convexity defined as in (46). As a result, the objective in (57) is $\frac{1}{2\eta}$-strongly convex according to the Sobolev norm $\|\cdot\|_{\dot{H}^1}$. In addition, this objective is minimized over a convex feasible set $\{D^{\boldsymbol{\theta}} : \theta \in \Theta\}$, due to the theorem's assumption. Therefore, Lemma 2 shows that the optimal $D^{\boldsymbol{\theta}^*}$ satisfies the following inequality for every $\boldsymbol{\theta}$:

$$\begin{aligned}&\frac{1}{4\eta} \mathbb{E}\big[\big\|\nabla D^{\boldsymbol{\theta}}(\mathbf{X})\big\|_2^2\big] - \frac{1}{4\eta} \mathbb{E}\big[\big\|\nabla D^{\boldsymbol{\theta}^*}(\mathbf{X})\big\|_2^2\big] \\ &\geq \frac{1}{4\eta} \big\| D^{\boldsymbol{\theta}} - D^{\boldsymbol{\theta}^*} \big\|_{\dot{H}^1}^2.\end{aligned}$$

The above result implies that

$$V(G_{\boldsymbol{\theta}^*}, D^{\boldsymbol{\theta}^*}) \leq V(G_{\boldsymbol{\theta}}, D^{\boldsymbol{\theta}}) - \frac{1}{4\eta} \big\| D^{\boldsymbol{\theta}} - D^{\boldsymbol{\theta}^*} \big\|_{\dot{H}^1}^2, \tag{58}$$

which completes the proof. □

3.7. Proof of Theorem 3

Theorem 3. *Consider the WGAN problem minimizing the first-order Wasserstein distance with cost function $c(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_2$. For each $G_{\boldsymbol{\theta}}$, let $\alpha_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}^{\geq 0}$ denote the magnitude of the resulted optimal transport map from $G_{\boldsymbol{\theta}}(\mathbf{Z})$ to $\mathbf{X}$, i.e. $\mathbf{X} - \alpha_{\boldsymbol{\theta}}^2(\mathbf{X})\nabla D^{\boldsymbol{\theta}}(\mathbf{X})$ shares the same distribution with $G_{\boldsymbol{\theta}}(\mathbf{Z})$. Given these definitions, assume that*

- *$\{\alpha_{\boldsymbol{\theta}}(\cdot)\nabla D^{\boldsymbol{\theta}}(\cdot) : \boldsymbol{\theta} \in \Theta\}$ is a convex set,*
- *for every $\mathbf{x}$ and $\boldsymbol{\theta}$ we have $\eta \leq \alpha_{\boldsymbol{\theta}}^2(\mathbf{x})$ for constant $\eta > 0$.*

Then, $(G_{\boldsymbol{\theta}^*}, D^{\boldsymbol{\theta}^*})$ for the Wasserstein distance-minimizing generator function $G_{\boldsymbol{\theta}^*}$ gives a η-proximal equilibrium with respect to the Sobolev norm in the main text.

Proof.

Lemma 3. *Consider two vectors $\mathbf{x}, \mathbf{y}$ with equal Euclidean norms $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2$. Then for every $0 \leq a \leq b$, we have*

$$a\|\mathbf{x} - \mathbf{y}\|_2 \leq \|a\mathbf{x} - b\mathbf{y}\|_2. \tag{59}$$

Proof. Note that

$$\begin{aligned}&\|a\mathbf{x} - b\mathbf{y}\|_2^2 - a^2\|\mathbf{x} - \mathbf{y}\|_2^2 \\ &= (b^2 - a^2)\|\mathbf{y}\|_2^2 - 2a(b - a)\mathbf{x}^T\mathbf{y}\end{aligned}$$

$$= (b-a)((b+a)\|\mathbf{y}\|_2^2 - 2a\mathbf{x}^T\mathbf{y})$$
$$\geq 0.$$

The above holds as we have assumed that $0 \leq a \leq b$ implying $0 \leq 2a \leq b+a$ and since the two vectors $\mathbf{x}, \mathbf{y}$ share the same Euclidean norm

$$|\mathbf{x}^T\mathbf{y}| \leq \frac{\|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2}{2} = \|\mathbf{y}\|_2^2.$$

Hence, Lemma 3's proof is complete. □

As shown by the Kantorovich duality (Villani, 2008), for the optimal $D^{\boldsymbol{\theta}}$ and the optimal coupling $\pi_{\boldsymbol{\theta}}(P_{\mathbf{X}}, P_{G_{\boldsymbol{\theta}}(\mathbf{Z})})$ the following holds with probability 1 for every joint sample $(\mathbf{X}, \mathbf{X}')$ drawn from the optimal coupling $\pi_{\boldsymbol{\theta}}$,

$$D^{\boldsymbol{\theta}}(\mathbf{X}) - D^{\boldsymbol{\theta}}(\mathbf{X}') = \|\mathbf{X} - \mathbf{X}'\|_2. \tag{60}$$

Knowing that $D^{\boldsymbol{\theta}}$ is 1-Lipschitz, for every convex combination $\beta\mathbf{X} + (1-\beta)\mathbf{X}'$ we must have

$$\nabla D^{\boldsymbol{\theta}}\big(\beta\mathbf{X} + (1-\beta)\mathbf{X}'\big) = \frac{\mathbf{X} - \mathbf{X}'}{\|\mathbf{X} - \mathbf{X}'\|}.$$

This will imply that there definitely exists $\alpha_{\boldsymbol{\theta}}$ such that the transport map described in the theorem maps the data distribution to the generative model. Plugging this transport map into the definition of the first-order Wasserstein distance, we obtain

$$\begin{aligned} V(G_{\boldsymbol{\theta}}, D^{\boldsymbol{\theta}}) :&= W_1(P_{\mathbf{X}}, P_{G_{\boldsymbol{\theta}}(\mathbf{Z})}) \\ &= \mathbb{E}\big[\big\|\alpha_{\boldsymbol{\theta}}^2(\mathbf{X})\nabla D^{\boldsymbol{\theta}}(\mathbf{X})\big\|_2\big] \\ &= \mathbb{E}\big[\big\|\alpha_{\boldsymbol{\theta}}(\mathbf{X})\nabla D^{\boldsymbol{\theta}}(\mathbf{X})\big\|_2^2\big] \end{aligned}$$

where the last equality holds since the Euclidean norm of $\nabla D^{\boldsymbol{\theta}}(\mathbf{X})$ has a unit Euclidean norm with probability 1 over the data distribution $P_{\mathbf{X}}$ as we proved $\nabla D^{\boldsymbol{\theta}}(\beta\mathbf{X} + (1-\beta)\mathbf{X}') = \frac{\mathbf{X}-\mathbf{X}'}{\|\mathbf{X}-\mathbf{X}'\|}$ holds for every $0 \leq \beta \leq 1$ including $\beta = 1$.

As a result, the Wasserstein GAN problem reduces to the following optimization problem

$$\min_{\boldsymbol{\theta}} \mathbb{E}\big[\big\|\alpha_{\boldsymbol{\theta}}(\mathbf{X})\nabla D^{\boldsymbol{\theta}}(\mathbf{X})\big\|_2^2\big] \tag{61}$$

Defining $h_{\boldsymbol{\theta}}(\mathbf{X}) := \alpha_{\boldsymbol{\theta}}(\mathbf{X})\nabla D^{\boldsymbol{\theta}}(\mathbf{X})$, $\frac{1}{2}\mathbb{E}\big[\|h_{\boldsymbol{\theta}}(\mathbf{X})\|_2^2\big]$ is 1-strongly convex with respect to the norm function

$$\|h\|_{\dot{H}^0} = \sqrt{\mathbb{E}\big[\,\|h(\mathbf{X})\|_2^2\,\big]}$$

that is induced by the following inner product and results in a Hilbert space

$$\langle D_1, D_2\rangle_{\dot{H}^0} := \mathbb{E}_{P_{\mathbf{X}}}[D_1(\mathbf{X})D_2(\mathbf{X})].$$

Therefore, for the $\boldsymbol{\theta}^*$ minimizing the objective in (61) over the assumed convex set $\{h_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, Lemma 2 implies that

$$\begin{aligned} &\mathbb{E}\big[\big\|\alpha_{\boldsymbol{\theta}}(\mathbf{X})\nabla D^{\boldsymbol{\theta}}(\mathbf{X})\big\|_2^2\big] - \mathbb{E}\big[\big\|\alpha_{\boldsymbol{\theta}^*}(\mathbf{X})\nabla D^{\boldsymbol{\theta}^*}(\mathbf{X})\big\|_2^2\big] \\ &= \mathbb{E}\big[\big\|h_{\boldsymbol{\theta}}(\mathbf{X})\big\|_2^2\big] - \mathbb{E}\big[\big\|h_{\boldsymbol{\theta}^*}(\mathbf{X})\big\|_2^2\big] \\ &\geq \|h_{\boldsymbol{\theta}} - h_{\boldsymbol{\theta}^*}\|_{\dot{H}^0}^2 \\ &= \mathbb{E}\big[\|h_{\boldsymbol{\theta}}(\mathbf{X}) - h_{\boldsymbol{\theta}^*}(\mathbf{X})\|_2^2\big] \\ &= \mathbb{E}\big[\|\alpha_{\boldsymbol{\theta}}(\mathbf{X})\nabla D^{\boldsymbol{\theta}}(\mathbf{X}) - \alpha_{\boldsymbol{\theta}^*}(\mathbf{X})\nabla D^{\boldsymbol{\theta}^*}(\mathbf{X})\|_2^2\big] \\ &\geq \eta\mathbb{E}\big[\|\nabla D^{\boldsymbol{\theta}}(\mathbf{X}) - \nabla D^{\boldsymbol{\theta}^*}(\mathbf{X})\|_2^2\big] \\ &= \eta\|D_{\boldsymbol{\theta}} - D_{\boldsymbol{\theta}^*}\|_{\dot{H}^1}^2. \end{aligned}$$

Here the last inequality follows from Lemma 3 since every $D^{\boldsymbol{\theta}}$ has a unit-norm gradient with probability 1 according to the data distribution $P_{\mathbf{X}}$. Therefore, we have proved that

$$V(G_{\boldsymbol{\theta}}, D^{\boldsymbol{\theta}}) - V(G_{\boldsymbol{\theta}^*}, D^{\boldsymbol{\theta}^*}) \geq \eta\|D_{\boldsymbol{\theta}} - D_{\boldsymbol{\theta}^*}\|_{\dot{H}^1}^2. \tag{62}$$

The above inequality results in the following for every feasible $\boldsymbol{\theta}$

$$V(G_{\boldsymbol{\theta}^*}, D^{\boldsymbol{\theta}^*}) \leq \max_{D\in\mathcal{D}} V(G_{\boldsymbol{\theta}}, D) - \eta\|D - D_{\boldsymbol{\theta}^*}\|_{\dot{H}^1}^2. \tag{63}$$

Hence, according to Proposition 2, we have shown that the pair $(G_{\boldsymbol{\theta}^*}, D^{\boldsymbol{\theta}^*})$ is an η-proximal equilibrium with respect to the Sobolev norm $\|\cdot\|_{\dot{H}^1}$. □

References

Ambrosio, L. and Gigli, N. A user's guide to optimal transport. In *Modelling and optimisation of flows on networks*, pp. 1–155. Springer, 2013.

Bernhard, P. and Rapaport, A. On a theorem of danskin with an application to a theorem of von neumann-sion. *Nonlinear Analysis: Theory, Methods & Applications*, 24 (8):1163–1181, 1995.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Feizi, S., Farnia, F., Ginart, T., and Tse, D. Understanding gans: the lqg setting. *arXiv preprint arXiv:1710.10793*, 2017.

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