A. Proofs

A.1. Proof of Theorem 1

Let X, Y be two random vectors such that $X \sim \mathcal{P}_g, Y \sim \mathcal{P}_r$. Assume $\mathbb{E}_{X\sim\mathcal{P}_g}\|X\| < \infty$ and $\mathbb{E}_{Y\sim\mathcal{P}_r}\|Y\| < \infty$. Let $\mathfrak{G}(f) = \mathbb{E}_{X\sim\mathcal{P}_g}\phi(f(X)) + \mathbb{E}_{Y\sim\mathcal{P}_r}\varphi(f(Y))$. Let $\|f\|_{Lip}$ denote the Lipschitz constant of f. Let $\mathcal{S}_r$ and $\mathcal{S}_g$ denote the supports of $\mathcal{P}_r$ and $\mathcal{P}_g$, respectively. Let $W_1(\mathcal{P}_r, \mathcal{P}_g)$ denote the 1-st Wasserstein distance between $\mathcal{P}_r$ and $\mathcal{P}_g$.

Lemma 1. *Let ϕ and φ be two convex functions, whose domains are both $\mathbb{R}$. Assume f is subject to $\|f\|_{Lip} \leq k$. If there is $a_0 \in \mathbb{R}$ such that $\phi'(a_0) + \varphi'(a_0) = 0$, then we have a lower bound for $\mathfrak{G}(f)$.*

Proof. Given that ϕ, φ are convex functions, we have

$$\begin{aligned}
\mathfrak{G}(f) &= \mathbb{E}_{X\sim\mathcal{P}_g}\phi(f(X)) + \mathbb{E}_{Y\sim\mathcal{P}_r}\varphi(f(Y)) \\
&\geq \mathbb{E}_{X\sim\mathcal{P}_g}(\phi'(a_0)(f(x)-a_0)+\phi(a_0)) + \mathbb{E}_{Y\sim\mathcal{P}_r}(\varphi'(a_0)(f(x)-a_0)+\varphi(a_0)) \\
&= \phi'(a_0)\mathbb{E}_{X\sim\mathcal{P}_g}f(x) + \varphi'(a_0)\mathbb{E}_{Y\sim\mathcal{P}_r}f(Y) + C \\
&= (\phi'(a_0)+\varphi'(a_0))\mathbb{E}_{X\sim\mathcal{P}_g}f(X) + \varphi'(a_0)(\mathbb{E}_{Y\sim\mathcal{P}_r}f(Y) - \mathbb{E}_{X\sim\mathcal{P}_g}f(X)) + C \\
&= k\varphi'(a_0)(\mathbb{E}_{Y\sim\mathcal{P}_r}\frac{1}{k}f(Y) - \mathbb{E}_{X\sim\mathcal{P}_g}\frac{1}{k}f(X)) + C \\
&\geq -k\varphi'(a_0)W_1(\mathcal{P}_r, \mathcal{P}_g) + C.
\end{aligned} \tag{13}$$

Therefore, we get the lower bound. □

Lemma 2. *Let ϕ and φ be two convex functions, whose domains are both $\mathbb{R}$. Assume f is subject to $\|f\|_{Lip} \leq k$.*

- *If there exists $a_1 \in \mathbb{R}$ such that $\phi'(a_1) + \varphi'(a_1) > 0$, then we have: if $f(0) \to +\infty$, then $\mathfrak{G}(f) \to +\infty$;*
- *If there exists $a_2 \in \mathbb{R}$ such that $\phi'(a_2) + \varphi'(a_2) < 0$, then we have: if $f(0) \to -\infty$, then $\mathfrak{G}(f) \to +\infty$.*

Proof. Since ϕ, φ are convex functions, we have

$$\begin{aligned}
\mathfrak{G}(f) &= \mathbb{E}_{X\sim\mathcal{P}_g}\phi(f(X)) + \mathbb{E}_{Y\sim\mathcal{P}_r}\varphi(f(Y)) \\
&\geq \mathbb{E}_{X\sim\mathcal{P}_g}(\phi'(a_1)(f(x)-a_1)+\phi(a_1)) + \mathbb{E}_{Y\sim\mathcal{P}_r}(\varphi'(a_1)(f(x)-a_1)+\varphi(a_1)) \\
&= \phi'(a_1)\mathbb{E}_{X\sim\mathcal{P}_g}f(x) + \varphi'(a_1)\mathbb{E}_{Y\sim\mathcal{P}_r}f(Y) + C_1 \\
&= (\phi'(a_1)+\varphi'(a_1))\mathbb{E}_{X\sim\mathcal{P}_g}f(X) + \varphi'(a_1)(\mathbb{E}_{Y\sim\mathcal{P}_r}f(Y) - \mathbb{E}_{X\sim\mathcal{P}_g}f(X)) + C_1 \\
&= (\phi'(a_1)+\varphi'(a_1))\mathbb{E}_{X\sim\mathcal{P}_g}f(X) + k\varphi'(a_1)(\mathbb{E}_{Y\sim\mathcal{P}_r}\frac{1}{k}f(Y) - \mathbb{E}_{X\sim\mathcal{P}_g}\frac{1}{k}f(X)) + C_1 \\
&\geq (\phi'(a_1)+\varphi'(a_1))\mathbb{E}_{X\sim\mathcal{P}_g}f(X) - k\varphi'(a_1)W_1(\mathcal{P}_r, \mathcal{P}_g) + C_1 \\
&\geq (\phi'(a_1)+\varphi'(a_1))f(0) - k(\phi'(a_1)+\varphi'(a_1))\mathbb{E}_{X\sim\mathcal{P}_g}\|X\| - k\varphi' W_1(\mathcal{P}_r, \mathcal{P}_g) + C_1.
\end{aligned} \tag{14}$$

Thus, if $f(0) \to +\infty$, then $\mathfrak{G}(f) \to +\infty$. And we can prove the other case symmetrically. □

Lemma 3. *Let ϕ and φ be two convex functions, whose domains are both $\mathbb{R}$. If ϕ and φ satisfy the following properties:*

- *$\phi' \geq 0, \varphi' \leq 0$;*
- *There exist $a_0, a_1, a_2 \in \mathbb{R}$ such that $\phi'(a_0) + \varphi'(a_0) = 0, \phi'(a_1) + \varphi'(a_1) > 0, \phi'(a_2) + \varphi'(a_2) < 0$.*

Then we have $\mathfrak{G}(f) = \mathbb{E}_{X\sim\mathcal{P}_r}\phi(f(X)) + \mathbb{E}_{Y\sim\mathcal{P}_g}\varphi(f(Y))$, where f is subject to $\|f\|_{Lip} \leq k$, has global minima. That is, $\exists f^*,\ s.t.$

- *$\|f^*\|_{Lip} \leq k$;*
- *$\forall f\ s.t.\ \|f\|_{Lip} \leq k$, we have $\mathfrak{G}(f^*) \leq \mathfrak{G}(f)$.*

Proof. According to Lemma 1, $\mathfrak{G}(f)$ has a lower bound, which means $inf(\mathfrak{G}(f)) > -\infty$. Thus we can get a series of functions $\{f_n\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty} \mathfrak{G}(f_n) = inf(\mathfrak{G}(f))$. Suppose that $\{r_i\}_{i=1}^{\infty}$ is the sequence of all rational points in $dom(f)$. Due to Lemma 2, for any $x \in \mathbb{R}$, $\{f_n(x)|n \in \mathbb{R}\}$ is bounded. By Bolzano-Weierstrass theorem, there is a subsequence $\{f_{1n}\} \subseteq \{f_n\}$ such that $\{f_{1n}(r_1)\}_{n=1}^{\infty}$ converges. And there is a subsequence $\{f_{2n}\} \subseteq \{f_{1n}\}$ such that $\{f_{2n}(r_2)\}_{n=1}^{\infty}$ converges. As for r_i, there is a subsequence $\{f_{in}\} \subseteq \{f_{i-1n}\}$ such that $\{f_{in}(r_i)\}_{n=1}^{\infty}$ converges. Then the sequence $\{f_{nn}\}_{n=1}^{\infty}$ will converge at r_i.

Furthermore, for all $x \in dom(f)$, we claim that $\{f_{nn}\}_{n=1}^{\infty}$ converges at x. Actually, $\forall \epsilon > 0$, find $r \in \{r_i\}$ such that $\|x - r\| \leq \frac{\epsilon}{10k}$, we have

$$\begin{aligned} \lim_{m,l\to\infty} |f_{mm}(x) - f_{ll}(x)| &\leq \lim_{m,l\to\infty} (|f_{mm}(x) - f_{mm}(r)| + |f_{mm}(r) - f_{ll}(r)| + |f_{ll}(r) - f_{ll}(x)|) \\ &\leq \lim_{m,l\to\infty} (\frac{\epsilon}{10} + \frac{\epsilon}{10} + |f_{mm}(r) - f_{ll}(r)|) = \frac{\epsilon}{5} \end{aligned} \tag{15}$$

Let $\epsilon \to 0$, then we get $\lim_{m,l\to\infty} |f_{mm}(x) - f_{ll}(x)| = 0$.

We denote $\{f_{nn}\}_{n=1}^{\infty}$ as $\{g_n\}_{n=1}^{\infty}$ and $\{g_n\}_{n=1}^{\infty}$ converges to g. Due to Lemma 2, we know that $\exists C'$ such that $|g_n(0)| \leq C', \ \forall n \in \mathbb{N}$. Because $\phi' \geq 0, \varphi' \leq 0$, we have

$$\phi(g_n(x)) \geq \phi(g_n(0) - k\|x\|) \geq \phi(-C' - k\|x\|) \geq \phi'(a_0)(-C' - k\|x\| - a_0) + \phi(a_0) = -k\phi'(a_0)\|x\| + C^{''} \tag{16}$$

That is, $\phi(g_n(x)) + k\phi'(a_0)\|x\| - C^{''} \geq 0$.

By Fatou's Lemma,

$$\begin{aligned} \mathbb{E}_{X\sim\mathcal{P}_g}(\phi(g(X)) + k\phi'(a_0)\|X\| - C^{''}) &= \mathbb{E}_{X\sim\mathcal{P}_g} \underline{\lim}_{n\to\infty} (\phi(g_n(X)) + k\phi'(a_0)\|X\| - C^{''}) \\ &\leq \underline{\lim}_{n\to\infty} \mathbb{E}_{X\sim\mathcal{P}_g}(\phi(g_n(X)) + k\phi'(a_0)\|X\| - C^{''}) \\ &= \underline{\lim}_{n\to\infty} \mathbb{E}_{X\sim\mathcal{P}_g}\phi(g_n(X)) + \mathbb{E}_{X\sim\mathcal{P}_g}(k\phi'(a_0)\|X\| - C^{''}) \end{aligned} \tag{17}$$

It means $\mathbb{E}_{X\sim\mathcal{P}_g}\phi(g(X)) \leq \underline{\lim}_{n\to\infty} \mathbb{E}_{X\sim\mathcal{P}_g}\phi(g_n(X))$. Similarly, we have $\mathbb{E}_{Y\sim\mathcal{P}_r}\varphi(g(Y)) \leq \underline{\lim}_{n\to\infty} \mathbb{E}_{Y\sim\mathcal{P}_r}\varphi(g_n(Y))$. Combining the two inequalities, we have

$$\begin{aligned} \mathfrak{G}(g) = \mathbb{E}_{X\sim\mathcal{P}_g}\phi(g(X)) + \mathbb{E}_{Y\sim\mathcal{P}_r}\varphi(g(Y)) &\leq \underline{\lim}_{n\to\infty} \mathbb{E}_{X\sim\mathcal{P}_g}\phi(g_n(X)) + \underline{\lim}_{n\to\infty} \mathbb{E}_{Y\sim\mathcal{P}_r}\varphi(g_n(Y)) \\ &\leq \underline{\lim}_{n\to\infty} (\mathbb{E}_{X\sim\mathcal{P}_g}\phi(g_n(X)) + \mathbb{E}_{Y\sim\mathcal{P}_r}\varphi(g_n(Y))) = \inf_{\|f\|_{Lip}\leq k} \mathfrak{G}(f) \end{aligned} \tag{18}$$

Note that for any $x, y \in dom(g)$, $|g(x) - g(y)| \leq \lim_{n\to\infty}(|g(x) - g_n(x)| + |g_n(x) - g_n(y)| + |g_n(y) - g(y)|) \leq k\|x - y\|$. That is, $\|g\|_{Lip} \leq k$, $\mathfrak{G}(g) = \inf_{\|f\|_{Lip}\leq k} \mathfrak{G}(f)$. □

Lemma 4 (Wasserstein distance)**.** *$\mathfrak{T}(f) = \mathbb{E}_{X\sim\mathcal{P}_g} f(X) - \mathbb{E}_{Y\sim\mathcal{P}_r} f(Y)$, where f is subject to $\|f\|_{Lip} \leq k$, has global minima.*

Proof. It is easy to find that for any $C \in \mathbb{R}$, $\mathfrak{T}(f + C) = \mathfrak{T}(f)$. Similar to the previous lemma, we can get a series of functions $\{f_n\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty} \mathfrak{T}(f_n) = inf(\mathfrak{T}(f))$. Without loss of generality, we assume that $f_n(0) = 0, \forall n \in \mathbb{N}^+$. Because $\|f_n\|_{Lip} \leq k$, we can claim that for any $x \in \mathbb{R}$, $\{f_n(x)|n \in \mathbb{R}\}$ is bounded. Then we can imitate the method used in Lemma 3 and find the optimal function f^* such that $\mathfrak{T}(f^*) = \inf_{\|f\|_{Lip}\leq k} \mathfrak{T}(f)$. □

Lemma 5. *Let ϕ and φ be two convex functions, whose domains are both $\mathbb{R}$. If we further suppose that the support sets $\mathcal{S}_r$ and $\mathcal{S}_g$ are bounded. Then if ϕ and φ satisfy the following properties:*

- *$\phi' \geq 0, \varphi' \leq 0$;*
- *There is $a_0 \in \mathbb{R}$ such that $\phi'(a_0) + \varphi'(a_0) = 0$.*

We have $\mathfrak{G}(f) = \mathbb{E}_{X\sim\mathcal{P}_g}\phi(f(X)) + \mathbb{E}_{Y\sim\mathcal{P}_r}\varphi(f(Y))$*, where* f *is subject to* $\|f\|_{Lip} \leq k$*, has global minima.*

That is, $\exists f^*,\ s.t.$

- $\|f^*\|_{Lip} \leq k$
- $\forall f\ s.t.\ \|f\|_{Lip} \leq k$ *, we have* $\mathfrak{G}(f^*) \leq \mathfrak{G}(f)$*.*

Proof. We have proved most conditions in previous lemmas. And we only have to consider the condition that for any $x \in \mathbb{R}$, $\phi'(x) + \varphi'(x) \geq 0$ (or $\phi'(x) + \varphi'(x) \leq 0$) and there exists a_1 such that $\phi'(a_1) + \varphi'(a_1) > 0$ (or $\phi'(a_1) + \varphi'(a_1) < 0$).

Without loss of generality, we assume that $\phi'(x) + \varphi'(x) \geq 0$ for all x and there exists a_1 such that $\phi'(a_1) + \varphi'(a_1) > 0$. Then we know $\forall x \leq a_0,\ \phi'(x) + \varphi'(x) = 0$, which leads to $\forall x \leq a_0,\ \phi'(x) = -\varphi'(x)$. Thus, for any $x \leq a_0$, $0 \leq \phi''(x) = -\varphi''(x) \leq 0$, which means $\forall x \leq a_0,\ \phi(x) = -\varphi(x) = tx,\ t \geq 0$. Similar to the previous lemmas, we can get a series of functions $\{f_n\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty} \mathfrak{G}(f_n) = inf(\mathfrak{G}(f))$. Actually we can assume that for all $n \in \mathbb{N}^+$, there is $f_n(0) \in [-C, C]$, where C is a constant. In fact, it is not difficult to find $f_n(0) \leq C$ with Lemma 2. On the other hand, when $C > k \cdot diam(\mathcal{S}_r \cup \mathcal{S}_g) + a_0$, then: if $f(0) < -C$, we have $f(X) < a_0$ for all $X \in \mathcal{S}_r \cup \mathcal{S}_g$. In this case, $\mathfrak{G}(f) = \mathfrak{G}(f - f(0) - C)$. This is the reason we can assume $f_n(0) \in [-C, C]$. Because $\|f_n\|_{Lip} \leq k$, we can assert that for any $x \in \mathbb{R}$, $\{f_n(x)|n \in \mathbb{R}\}$ is bounded. So we can imitate the method used in Lemma 3 and find the optimal function f^* such that $\mathfrak{G}(f^*) = \inf_{\|f\|_{Lip}\leq k} \mathfrak{G}(f)$. □

Lemma 6 (Theorem 1 Part I)**.** *Under the same assumption of Lemma 5, we have* $\mathfrak{F}(f) = \mathbb{E}_{X\sim\mathcal{P}_g}\phi(f(X)) + \mathbb{E}_{Y\sim\mathcal{P}_r}\varphi(f(Y)) + \lambda\|f\|_{Lip}^{\alpha}$ *with* $\lambda > 0$ *and* $\alpha > 1$ *has global minima.*

Proof. When $\|f\|_{Lip} = \infty$, it is trivial that $\mathfrak{F}(f) = \infty$. And when $\|f\|_{Lip} < \infty$, combining Lemma 1, we have $\mathfrak{F}(f) = \mathfrak{G}(f) + \lambda\|f\|_{Lip}^{\alpha} \geq -\|f\|_{Lip}\varphi'(a_0)W_1(\mathcal{P}_r, \mathcal{P}_g) + \lambda\|f\|_{Lip}^{\alpha}$. When $\lambda > 0$ and $\alpha > 1$, the right term is a convex function about $\|f\|_{Lip}$, it has a lower bound. So we can find a sequence $\{f_n\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty} \mathfrak{F}(f_n) = \inf_{f\in dom} \mathfrak{F}(f)$. It is no doubt that there exists a constant C such that $\|f_n\|_{Lip} \leq C$ for all f_n. Then it is not difficult to show for any point x, $\{f_n(x)\}$ is bounded. So we can imitate the method used in main theorem to find the sequence $\{g_n\}$ such that $\{g_n\} \subseteq \{f_n\}$ and $\{g_n\}_{n=1}^{\infty}$ converge at every point x. Suppose $\lim_{n\to\infty} g_n = g$, then by Fatou's Lemma, we have $\mathfrak{G}(g) \leq \underline{\lim}_{n\to\infty} \mathfrak{G}(g_n)$.

Next, We prove that $\|g\|_{Lip} \leq \underline{\lim}_{n\to\infty} \|g_n\|_{Lip}$. If the claim holds, then $\mathfrak{F}(g) = \mathfrak{G}(g) + \lambda\|g\|_{Lip}^{\alpha} \leq \underline{\lim}_{n\to\infty} \mathfrak{G}(g_n) + \underline{\lim}_{n\to\infty} \lambda\|g_n\|_{Lip}^{\alpha} \leq \underline{\lim}_{n\to\infty}(\mathfrak{G}(g_n) + \lambda\|g_n\|_{Lip}^{\alpha}) = \inf \mathfrak{F}(f)$. Thus, the global minima exists. In fact, if $\|g\|_{Lip} > \underline{\lim}_{n\to\infty} \|g_n\|_{Lip}$, then there exist x, y such that $\frac{|g(x)-g(y)|}{\|x-y\|} \geq \underline{\lim}_{n\to\infty} \|g_n\|_{Lip} + \epsilon \geq \underline{\lim}_{n\to\infty} \frac{|g_n(x)-g_n(y)|}{\|x-y\|} + \epsilon$. i.e. $|g(x) - g(y)| \geq \underline{\lim}_{n\to\infty} |g_n(x) - g_n(y)| + \epsilon\|x - y\| = |g(x) - g(y)| + \epsilon\|x - y\| > |g(x) - g(y)|$. The contradiction tells us that $\|g\|_{Lip} \leq \underline{\lim}_{n\to\infty} \|g_n\|_{Lip}$. □

Lemma 7 (Theorem 1 Part II)**.** *Let* ϕ *and* φ *be two convex functions, whose domains are both* $\mathbb{R}$*. If* ϕ *or* φ *is strictly convex, then the minimizer of* $\mathfrak{F}(f) = \mathbb{E}_{X\sim\mathcal{P}_g}\phi(f(X)) + \mathbb{E}_{Y\sim\mathcal{P}_r}\varphi(f(Y)) + \lambda\|f\|_{Lip}^{\alpha}$ *with* $\lambda > 0$ *and* $\alpha > 1$ *is unique (in the support of* $\mathcal{S}_r \cup \mathcal{S}_g$*).*

Proof. Without loss of generality, we assume that ϕ is strictly convex. By the strict convexity of ϕ, we have $\forall x, y \in \mathbb{R},\ \phi(\frac{x+y}{2}) < \frac{1}{2}(\phi(x) + \phi(y))$. Assume f_1 and f_2 are two different minimizers of $\mathfrak{F}(f)$.

First, we have

$$\begin{aligned}
\left\|\frac{f_1 + f_2}{2}\right\|_{Lip} &= \sup_{x,y} \frac{\frac{f_1(x)+f_2(x)}{2} - \frac{f_1(y)+f_2(y)}{2}}{\|x - y\|} \\
&\leq \sup_{x,y} \frac{1}{2}\frac{|f_1(x) - f_1(y)| + |f_2(x) - f_2(y)|}{\|x - y\|} \\
&\leq \frac{1}{2}\Big(\sup_{x,y} \frac{|f_1(x) - f_1(y)|}{\|x - y\|} + \sup_{x,y} \frac{|f_2(x) - f_2(y)|}{\|x - y\|}\Big) \\
&= \frac{1}{2}(\|f_1\|_{Lip} + \|f_2\|_{Lip}).
\end{aligned} \tag{19}$$

And given $\lambda > 0$ and $\alpha > 1$, we further have

$$\begin{aligned}\lambda\Big\|\frac{f_1+f_2}{2}\Big\|_{Lip}^{\alpha} &\leq \lambda\Big(\frac{1}{2}(\|f_1\|_{Lip}+\|f_2\|_{Lip})\Big)^{\alpha} \\ &\leq \lambda\frac{1}{2}(\|f_1\|_{Lip}^{\alpha}+\|f_2\|_{Lip}^{\alpha}).\end{aligned} \tag{20}$$

Let $\mathfrak{F}(f_1) = \mathfrak{F}(f_2) = \inf \mathfrak{F}(f)$. Then we have

$$\begin{aligned}\mathfrak{G}\Big(\frac{f_1+f_2}{2}\Big) &= \mathbb{E}_{X\sim\mathcal{P}_g}\phi\Big(\frac{f_1+f_2}{2}\Big) + \mathbb{E}_{Y\sim\mathcal{P}_r}\varphi\Big(\frac{f_1+f_2}{2}\Big) + \lambda\Big\|\frac{f_1+f_2}{2}\Big\|_{Lip}^{\alpha} \\ &< \mathbb{E}_{X\sim\mathcal{P}_g}\Big(\frac{\phi(f_1)+\phi(f_2)}{2}\Big) + \mathbb{E}_{Y\sim\mathcal{P}_r}\varphi\Big(\frac{f_1+f_2}{2}\Big) + \lambda\Big\|\frac{f_1+f_2}{2}\Big\|_{Lip}^{\alpha} \\ &\leq \mathbb{E}_{X\sim\mathcal{P}_g}\Big(\frac{\phi(f_1)+\phi(f_2)}{2}\Big) + \mathbb{E}_{Y\sim\mathcal{P}_r}\Big(\frac{\varphi(f_1)+\varphi(f_2)}{2}\Big) + \lambda\Big\|\frac{f_1+f_2}{2}\Big\|_{Lip}^{\alpha} \\ &\leq \mathbb{E}_{X\sim\mathcal{P}_g}\Big(\frac{\phi(f_1)+\phi(f_2)}{2}\Big) + \mathbb{E}_{Y\sim\mathcal{P}_r}\Big(\frac{\varphi(f_1)+\varphi(f_2)}{2}\Big) + \lambda\frac{1}{2}(\|f_1\|_{Lip}^{\alpha}+\|f_2\|_{Lip}^{\alpha}) \\ &= \frac{1}{2}(\mathfrak{G}(f_1)+\mathfrak{G}(f_2)) = \inf \mathfrak{G}(f)\end{aligned} \tag{21}$$

We get a contradiction $\mathfrak{G}(\frac{f_1+f_2}{2}) < \inf \mathfrak{G}(f)$, which implies that the minimizer of $\mathfrak{G}(f)$ is unique. □

A.2. Proof of Theorem 2

Let $J_D = \mathbb{E}_{x\sim\mathcal{P}_g}[\phi(f(x))]+\mathbb{E}_{x\sim\mathcal{P}_r}[\varphi(f(x))]$. Let $\mathring{J}_D(x) = \mathcal{P}_g(x)\phi(f(x))+\mathcal{P}_r(x)\varphi(f(x))$. Clearly, $J_D = \int_{\mathbb{R}^n} \mathring{J}_D(x)dx$. Let $J_D^*(k) = \min_{f\in\mathcal{F}_{\textbf{k}\text{-Lip}}} J_D = \min_{f\in\mathcal{F}_{\text{1-Lip}},b} \mathbb{E}_{x\sim\mathcal{P}_g}[\phi(k\cdot f(x)+b)] + \mathbb{E}_{x\sim\mathcal{P}_r}[\varphi(k\cdot f(x)+b)]$.

Let $k(f)$ denote the Lipschitz constant of f. Define $J = J_D + \lambda\cdot k(f)^2$ and $f^* = \arg\min_f[J_D + \lambda\cdot k(f)^2]$.

Lemma 8. *It holds* $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} = 0$ *for all* x*, if and only if,* $k(f^*) = 0$.

Proof.

(i) If $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} = 0$ holds for all x, then $k(f^*) = 0$.

For the optimal f^*, it holds that $\frac{\partial J}{\partial k(f^*)} = \frac{\partial J_D^*}{\partial k(f^*)} + 2\lambda\cdot k(f^*) = 0$.

$\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} = 0$ for all x implies $\frac{\partial J_D^*}{\partial k(f^*)} = 0$. Thus we conclude that $k(f^*) = 0$.

(ii) If $k(f^*) = 0$, then $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} = 0$ holds for all x.

For the optimal f^*, it holds that $\frac{\partial J}{\partial k(f^*)} = \frac{\partial J_D^*}{\partial k(f^*)} + 2\lambda\cdot k(f^*) = 0$.

$k(f^*) = 0$ implies $\frac{\partial J_D^*}{\partial k(f^*)} = 0$. $k(f^*) = 0$ also implies $\forall x, y, f^*(x) = f^*(y)$.

Given $\forall x, y, f^*(x) = f^*(y)$, if there exists some point x such that $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} \neq 0$, then it is obvious that $\frac{\partial J_D^*}{\partial k(f^*)} \neq 0$.

It is contradictory to $\frac{\partial J_D^*}{\partial k(f^*)} = 0$. Thus we have $\forall x$, $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} = 0$. □

Lemma 9. *If* $\forall x, y, f^*(x) = f^*(y)$*, then* $\mathcal{P}_r = \mathcal{P}_g$.

Proof. $\forall x, y, f^*(x) = f^*(y)$ implies $k(f^*) = 0$. According to Lemma 8, for all x it holds $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} = 0$, i.e., $\mathcal{P}_g(x)\frac{\partial\phi(f^*(x))}{\partial f^*(x)}$ $+\mathcal{P}_r(x)\frac{\partial\varphi(f^*(x))}{\partial f^*(x)} = 0$. Thus, $\frac{\mathcal{P}_g(x)}{\mathcal{P}_r(x)} = -\frac{\frac{\partial\varphi(f^*(x))}{\partial f^*(x)}}{\frac{\partial\phi(f^*(x))}{\partial f^*(x)}}$. That is, $\frac{\mathcal{P}_g(x)}{\mathcal{P}_r(x)}$ has a constant value, which straightforwardly implies $\mathcal{P}_r = \mathcal{P}_g$. □

Proof of Theorem 2.

(a): Let k be the Lipschitz constant of f^*. Consider x with $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} \neq 0$. Define $k(x) = \sup_y \frac{|f(y)-f(x)|}{\|y-x\|}$.

(i) If $\forall \delta$ s.t. $\forall \epsilon$ there exist $z, w \in B(x, \epsilon)$ such that $\frac{|f^*(z)-f^*(w)|}{\|z-w\|} \geq k - \delta$, which means there exists t such that $f'(t) \geq k - \delta$, because $\frac{|f^*(z)-f^*(w)|}{\|z-w\|} = \frac{\int_w^z f^{*\prime}(t)dt}{\|z-w\|}$. Let $\epsilon \to 0$, we have $t \to x$. Then $|f^{*\prime}(t)| \to |f^{*\prime}(x)|$. Let $\delta \to 0$, we have $(k - \delta) \to k$. Assume f^* is smooth, we have that $|f'(x)| = k$, which means there exists a y such that $|f^*(y) - f^*(x)| = k\|y - x\|$.

(ii) Assume that $\exists \delta$ s.t. $\exists \epsilon$ and for all $z, w \in B(x, \epsilon)$, $\frac{|f^*(z)-f^*(w)|}{\|z-w\|} < k - \delta$. Consider the following condition, for all δ_2 and $\epsilon_2 \in (0, \epsilon/2)$, $\exists y \in B(x, \epsilon_2)$, such that $k(y) > k - \delta_2$. Then there exists a sequence of $\{y_n\}_{n=1}^{\infty}$ s.t. $\lim_{n\to\infty} \frac{|f(y)-f(y_n)|}{\|y-y_n\|} = k(y)$. Then there exists a y' such that $\frac{|f(y)-f(y')|}{\|y-y'\|} \geq k - \delta_2$. According to the assumption, we have $\|y - y'\| \geq \frac{\epsilon}{2}$. Then $k(x) \geq \frac{|f^*(x)-f^*(y)|}{\|x-y\|} \geq \frac{|f^*(y)-f^*(y')|-|f^*(x)-f^*(y)|}{\|x-y\|+\|y-y'\|} \geq \frac{|f^*(y)-f^*(y')|-k\|x-y\|}{\|x-y\|+\|y-y'\|} \geq (k - \delta_2)\frac{\|y-y'\|}{\|x-y\|+\|y-y'\|} - k\frac{\|x-y\|}{\|x-y\|+\|y-y'\|} \geq (1 - \frac{\epsilon_2}{\epsilon_2+\|y-y'\|})(k - \delta_2) - k\frac{\epsilon_2}{\|y-y'\|} \geq (1 - \frac{\epsilon_2}{\epsilon_2+\|y-y'\|})(k - \delta_2) - k\frac{\epsilon_2}{\|y-y'\|}$. Let $\epsilon_2 \to 0$ and $\delta_2 \to 0$. We get $k(x) = k$, which means there exists a y such that $|f^*(y) - f^*(x)| = k\|y - x\|$.

(iii) Now we can assume $\exists \delta_2$ s.t. $\exists \epsilon_2$ and for all $y \in B(x, \epsilon_2)$, such that $k(y) \leq k - \delta_2$. If $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} \neq 0$, without loss of generality, we can assume $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} > 0$. Then, for all $y \in B(x, \epsilon_2)$, we have $\frac{\partial \mathring{J}_D(y)}{\partial f^*(y)} > 0$, as long as ϵ_2 is small enough. Now we change the value of $f^*(y)$ for $y \in B(x, \epsilon_2)$. Let $g(y) = \begin{cases} f^*(y) - \frac{\epsilon_2}{N}(1 - \frac{\|x-y\|}{\epsilon_2}), & y \in B(x, \epsilon_2); \\ f^*(y) & \text{otherwise.} \end{cases}$. Because $\frac{\partial \mathring{J}_D(y)}{\partial f^*(y)} > 0, \forall y \in B(x, \epsilon_2)$, when N is sufficiently large, it is not difficult to show $J_D(g) < J_D(f^*)$. We next verify that $\|g\|_{Lip} \leq k$. For any y, z, if $y, z \notin B(x, \epsilon_2)$, then $\frac{|g(y)-g(z)|}{\|y-z\|} = \frac{|f^*(y)-f^*(z)|}{\|y-z\|} < k$. If $y \in B(x, \epsilon_2)$, $z \notin B(x, \epsilon_2)$, then $\frac{|g(y)-g(z)|}{\|y-z\|} \leq \frac{|(f^*(y)-f^*(z)|+\frac{\epsilon_2}{N}(1-\frac{\|x-y\|}{\epsilon_2}))}{\|y-z\|} \leq \frac{|f^*(y)-f^*(z)|}{\|y-z\|} + \frac{\frac{\epsilon_2}{N}(1-\frac{\|x-y\|}{\epsilon_2})}{\epsilon_2-\|x-y\|} = \frac{|(f^*(y)-f^*(z)|}{\|y-z\|} + \frac{1}{N} \leq k(y) + \frac{1}{N} \leq k - \delta_2 + \frac{1}{N} < k$ (when $N \gg \frac{1}{\delta_2}$). If $y, z \in B(x, \epsilon)$, then $\frac{|g(y)-g(z)|}{\|y-z\|} \leq \frac{|f^*(y)-f^*(z)|+|\frac{\epsilon_2}{N}(1-\frac{\|x-y\|}{\epsilon_2})-\frac{\epsilon_2}{N}(1-\frac{\|x-z\|}{\epsilon_2})|}{\|y-z\|} = \frac{|f^*(y)-f^*(z)|}{\|y-z\|} + \frac{\frac{\epsilon_2}{N}(\frac{\|x-y\|-\|x-z\|}{\epsilon_2})|}{\|y-z\|} \leq \frac{|f^*(y)-f^*(z)|}{\|y-z\|} + \frac{1}{N}\frac{\|y-z\|}{\|y-z\|} = \frac{|f^*(y)-f^*(z)|}{\|y-z\|} + \frac{1}{N} \leq k - \delta_2 + \frac{1}{N} < k$ (when $N \gg \frac{1}{\delta_2}$). So, we have $\|g\|_{Lip} \leq k$. But we have $J_D(g) < J_D(f^*)$. The contradiction tells us that there must exists a y such that $|f^*(y) - f^*(x)| = k\|y - x\|$.

(b): For $x \in \mathcal{S}_r \cup \mathcal{S}_g - \mathcal{S}_r \cap \mathcal{S}_g$, assuming $\mathcal{P}_g(x) \neq 0$ and $\mathcal{P}_r(x) = 0$, we have $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} = \mathcal{P}_g(x)\frac{\partial \phi(f^*(x))}{\partial f^*(x)} + \mathcal{P}_r(x)\frac{\partial \varphi(f^*(x))}{\partial f^*(x)} = \mathcal{P}_g(x)\frac{\partial \phi(f^*(x))}{\partial f^*(x)} > 0$, because $\mathcal{P}_g(x) > 0$ and $\frac{\partial \phi(f^*(x))}{\partial f^*(x)} > 0$. Then according to (a), there must exist a y such that $|f^*(y) - f^*(x)| = k(f^*) \cdot \|y - x\|$. The other situation can be proved in the same way.

(c): According to Lemma 9, in the situation that $\mathcal{P}_r \neq \mathcal{P}_g$, for the optimal f^*, there must exist at least one pair of points x and y such that $y \neq x$ and $f^*(x) \neq f^*(y)$. It also implies that $k(f^*) > 0$. Then according to Lemma 8, there exists a point x such that $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} \neq 0$. According to (a), there exists y with $y \neq x$ satisfying that $|f^*(y) - f^*(x)| = k(f^*) \cdot \|y - x\|$.

(d): In Nash equilibrium state, it holds that, for any $x \in \mathcal{S}_r \cup \mathcal{S}_g$, $\frac{\partial J}{\partial k(f)} = \frac{\partial J_D^*}{\partial k(f)} + 2\lambda \cdot k(f) = 0$ and $\frac{\partial \mathring{J}_D(x)}{\partial f(x)}\frac{\partial f(x)}{\partial x} = 0$. We claim that in the Nash equilibrium state, the Lipschitz constant $k(f)$ must be 0. If $k(f) \neq 0$, according to Lemma 8, there must exist a point $\hat{x}$ such that $\frac{\partial \mathring{J}_D(\hat{x})}{\partial f(\hat{x})} \neq 0$. And according to (a), it must hold that $\exists \hat{y}$ fitting $|f(\hat{y}) - f(\hat{x})| = k(f) \cdot \|\hat{x} - \hat{y}\|$. According to Theorem 4, we have $\|\frac{\partial f(\hat{x})}{\partial \hat{x}}\| = k(f) \neq 0$. This is contradictory to that $\frac{\partial \mathring{J}_D(\hat{x})}{\partial f(\hat{x})}\frac{\partial f(\hat{x})}{\partial \hat{x}} = 0$. Thus $k(f) = 0$. That is, $\forall x \in \mathcal{S}_r \cup \mathcal{S}_g$, $\frac{\partial f(x)}{\partial x} = 0$, which means $\forall x, y, f(x) = f(y)$. According to Lemma 9, $\forall x, y, f(x) = f(y)$ implies $\mathcal{P}_r = \mathcal{P}_g$. Thus $\mathcal{P}_r = \mathcal{P}_g$ is the only Nash equilibrium in our system. □

Remark 1. *For the Wasserstein distance, $\nabla_{f^*(x)}\mathring{J}_D(x) = 0$ if and only if $\mathcal{P}_r(x) = \mathcal{P}_g(x)$. For the Wasserstein distance, penalizing the Lipschitz constant also benefits: at the convergence state, it will hold $\frac{\partial f^*(x)}{\partial x} = 0$ for all x.*

A.3. Proof of Theorem 3

Lemma 10. *Let k be the Lipschitz constant of f. If $f(a) - f(b) = k\|a - b\|$ and $f(b) - f(c) = k\|b - c\|$, then $f(a) - f(c) = k\|a - c\|$ and $(a, f(a)), (b, f(b)), (c, f(c))$ lies in the same line.*

Proof. $f(a) - f(c) = f(a) - f(b) + f(b) - f(c) = k\|a - b\| + k\|b - c\| \geq k\|a - c\|$. Because the Lipschitz constant of f is k, we have $f(a) - f(c) \leq k\|a - c\|$. Thus $f(a) - f(c) = k\|a - c\|$. Because the triangle equality holds, we have a, b, c is in the same line. Furthermore, because $f(a) - f(b) = k\|a - b\|$, $f(b) - f(c) = k\|b - c\|$ and $f(a) - f(c) = k\|a - c\|$, we have $(a, f(a)), (b, f(b)), (c, f(c))$ lies in the same line. □

Lemma 11. *For any x with $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} > 0$, there exists a y with $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} < 0$ such that $f^*(y) - f^*(x) = k(f^*)\|y - x\|$.*

For any y with $\frac{\partial \mathring{J}_D(y)}{\partial f^*(y)} < 0$, there exists a x with $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} > 0$ such that $f^*(y) - f^*(x) = k(f^*)\|y - x\|$.

Proof. Consider x with $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} > 0$. According to Theorem 2, there exists y such that $|f^*(y) - f^*(x)| = k(f^*)\|y - x\|$. Assume that for every y that holds $|f^*(y) - f^*(x)| = k(f^*)\|y - x\|$, it has $\frac{\partial \mathring{J}_D(y)}{\partial f^*(y)} \geq 0$. Consider the set $S(x) = \{y \mid f^*(y) - f^*(x) = k(f^*)\|y - x\|\}$. Note that, according to Lemma 10, any z that holds $f^*(z) - f^*(y) = k(f^*)\|z - y\|$ for any $y \in S(x)$ will also be in $S(x)$. Similar as the proof of (a) in Theorem 2, we can decrease the value of $f^*(y)$ for all $y \in S(x)$ to construct a better f. By contradiction, we have that there must exist a y with $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} < 0$ such that $|f^*(y) - f^*(x)| = k(f^*)\|y - x\|$. Given the fact $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} > 0$ and $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} < 0$, we can conclude that $f^*(y) > f^*(x)$ and $f^*(y) - f^*(x) = k(f^*)\|y - x\|$. Otherwise, if $f^*(x) - f^*(y) = k(f^*)\|y - x\|$, then we can construct a better f by decreasing $f^*(x)$ and increasing $f^*(y)$ which does not break the k-Lipschitz constraint. The other case can be proved similarly. □

Lemma 12. *For any x, if $\frac{\partial \mathring{J}_D(x)}{\partial f(x)} > 0$, then $\mathcal{P}_g(x) > 0$. For any y, if $\frac{\partial \mathring{J}_D(y)}{\partial f(y)} < 0$, then $\mathcal{P}_r(y) > 0$.*

Proof. $\frac{\partial \mathring{J}_D(x)}{\partial f(x)} = \mathcal{P}_g(x)\frac{\partial \phi(f(x))}{\partial f(x)} + \mathcal{P}_r(x)\frac{\partial \varphi(f(x))}{\partial f(x)}$. And we know $\phi'(x) > 0$ and $\varphi'(x) < 0$. Naturally, $\frac{\partial \mathring{J}_D(x)}{\partial f(x)} > 0$ implies $\mathcal{P}_g(x) > 0$. Similarly, $\frac{\partial \mathring{J}_D(y)}{\partial f(y)} < 0$ implies $\mathcal{P}_r(y) > 0$. □

Proof of Theorem 3.

For any $x \in \mathcal{S}_g$, if $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} > 0$, according to Lemma 11, there exists a y with $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} < 0$ such that $f^*(y) - f^*(x) = k(f^*)\|y - x\|$. According to Lemma 12, we have $\mathcal{P}_r(y) > 0$. That is, there is a $y \in \mathcal{S}_r$ such that $f^*(y) - f^*(x) = k(f^*)\|y - x\|$. We can prove the other case symmetrically. □

Remark 2. $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} < 0$ *for some $x \in \mathcal{S}_g$ means x is at the overlapping region of $\mathcal{S}_r$ and $\mathcal{S}_g$. It can be regarded as a $y \in \mathcal{S}_r$, and one can apply the other rule which guarantees that there exists a $x' \in \mathcal{S}_g$ that bounds this point.*

A.4. Proof of Theorem 4

In this section, we will prove Theorem 4, i.e., Lipschitz continuity with l_2-norm (Euclidean Distance) can guarantee that the gradient is directly pointing towards some sample.

Let (x, y) be such that $y \neq x$, and we define $x_t = x + t \cdot (y - x)$ with $t \in [0, 1]$.

Lemma 13. *If $f(x)$ is k-Lipschitz with respect to $\|.\|_p$ and $f(y) - f(x) = k\|y - x\|_p$, then $f(x_t) = f(x) + t \cdot k\|y - x\|_p$*

Proof. As we know $f(x)$ is k-Lipschitz, with the property of norms, we have

$$\begin{aligned} f(y) - f(x) &= f(y) - f(x_t) + f(x_t) - f(x) \\ &\leq f(y) - f(x_t) + k\|x_t - x\|_p = f(y) - f(x_t) + t \cdot k\|y - x\|_p \\ &\leq k\|y - x_t\|_p + t \cdot k\|y - x\|_p = k \cdot (1 - t)\|y - x\|_p + t \cdot k\|y - x\|_p \\ &= k\|y - x\|_p. \end{aligned} \tag{22}$$

$f(y) - f(x) = k\|y-x\|_p$ implies all the inequalities is equalities. Therefore, $f(x_t) = f(x) + t \cdot k\|y-x\|_p$. □

Lemma 14. *Let v be the unit vector $\frac{y-x}{\|y-x\|_2}$. If $f(x_t) = f(x) + t \cdot k\|y-x\|_2$, then $\frac{\partial f(x_t)}{\partial v}$ equals to k.*

Proof.

$$\begin{aligned}\frac{\partial f(x_t)}{\partial v} &= \lim_{h\to 0}\frac{f(x_t + hv) - f(x_t)}{h} = \lim_{h\to 0}\frac{f(x_t + h\frac{y-x}{\|y-x\|_2}) - f(x_t)}{h}\\ &= \lim_{h\to 0}\frac{f(x_{t+\frac{h}{\|y-x\|_2}}) - f(x_t)}{h} = \lim_{h\to 0}\frac{\frac{h}{\|y-x\|_2}\cdot k\|y-x\|_2}{h} = k. \quad \square\end{aligned}$$

Proof of Theorem 4. Assume $p = 2$. According to (Adler & Lunz, 2018), if $f(x)$ is k-Lipschitz with respect to $\|.\|_2$ and $f(x)$ is differentiable at x_t, then $\|\nabla f(x_t)\|_2 \le k$. Let v be the unit vector $\frac{y-x}{\|y-x\|_2}$. We have

$$k^2 = k\frac{\partial f(x_t)}{\partial v} = k\langle v, \nabla f(x_t)\rangle = \langle kv, \nabla f(x_t)\rangle \le \|kv\|_2\|\nabla f(x_t)\|_2 = k^2. \tag{23}$$

Because the equality holds only when $\nabla f(x_t) = kv = k\frac{y-x}{\|y-x\|_2}$, we have that $\nabla f(x_t) = k\frac{y-x}{\|y-x\|_2}$. □

A.5. Proof of the New Dual Form of Wasserstein Distance

We here provide a proof for our new dual form of Wasserstein distance, i.e., Eq. (4).

The Wasserstein distance is given as follows

$$W_1(\mathcal{P}_r, \mathcal{P}_g) = \inf_{\pi\in\Pi(\mathcal{P}_r,\mathcal{P}_g)} \mathbb{E}_{(x,y)\sim\pi}\,[d(x,y)], \tag{24}$$

where $\Pi(\mathcal{P}_r, \mathcal{P}_g)$ denotes the set of all probability measures with marginals $\mathcal{P}_r$ and $\mathcal{P}_g$ on the first and second factors, respectively. The Kantorovich-Rubinstein (KR) dual (Villani, 2008) is written as

$$\begin{aligned} W_{KR}(\mathcal{P}_r, \mathcal{P}_g) = \sup\nolimits_f\ \mathbb{E}_{x\sim\mathcal{P}_r}[f(x)] - \mathbb{E}_{x\sim\mathcal{P}_g}[f(x)],\\ \textit{s.t. } f(x) - f(y) \le d(x,y),\ \forall x, \forall y.\end{aligned} \tag{25}$$

We will prove that Wasserstein distance in its dual form can also be written as

$$\begin{aligned} W_{LL}(\mathcal{P}_r, \mathcal{P}_g) = \sup\nolimits_f\ \mathbb{E}_{x\sim\mathcal{P}_r}[f(x)] - \mathbb{E}_{x\sim\mathcal{P}_g}[f(x)],\\ \textit{s.t. } f(x) - f(y) \le d(x,y),\ \forall x\in\mathcal{S}_r, \forall y\in\mathcal{S}_g,\end{aligned} \tag{26}$$

which relaxes the constraint in the KR dual form of Wasserstein distance.

Theorem 5. *Given $W_{KR}(\mathcal{P}_r, \mathcal{P}_g) = W_1(\mathcal{P}_r, \mathcal{P}_g)$, we have $W_{KR}(\mathcal{P}_r, \mathcal{P}_g) = W_{LL}(\mathcal{P}_r, \mathcal{P}_g) = W_1(\mathcal{P}_r, \mathcal{P}_g)$.*

Proof.

(i) For any f that satisfies "$f(x) - f(y) \le d(x,y),\ \forall x, \forall y$", it must satisfy "$f(x) - f(y) \le d(x,y),\ \forall x\in\mathcal{S}_r, \forall y\in\mathcal{S}_g$".

Thus, $W_{KR}(\mathcal{P}_r, \mathcal{P}_g) \le W_{LL}(\mathcal{P}_r, \mathcal{P}_g)$.

(ii) Let $F_{LL} = \{f |\, f(x) - f(y) \le d(x,y),\ \forall x\in\mathcal{S}_r, \forall y\in\mathcal{S}_g\}$.

Let $A = \{(x,y)\,|\,x\in\mathcal{S}_r, y\in\mathcal{S}_g\}$ and $I_A = \begin{cases}1, & (x,y)\in A;\\ 0, & otherwise\end{cases}$.

Let A^c denote the complementary set of A and define I_{A^c} accordingly.

$\forall\pi\in\Pi(\mathcal{P}_r, \mathcal{P}_g)$, we have the following:

$$
\begin{aligned}
W_{LL}(\mathcal{P}_r, \mathcal{P}_g) &= \sup\nolimits_{f \in F_{LL}} \ \mathbb{E}_{x \sim \mathcal{P}_r}\,[f(x)] - \mathbb{E}_{x \sim \mathcal{P}_g}\,[f(x)] \\
&= \sup\nolimits_{f \in F_{LL}} \ \mathbb{E}_{(x,y) \sim \pi}[f(x) - f(y)] \\
&= \sup\nolimits_{f \in F_{LL}} \ \mathbb{E}_{(x,y) \sim \pi}[(f(x) - f(y)) I_A] + \mathbb{E}_{(x,y) \sim \pi}[(f(x) - f(y)) I_{A^c}] \\
&= \sup\nolimits_{f \in F_{LL}} \ \mathbb{E}_{(x,y) \sim \pi}[(f(x) - f(y)) I_A] \\
&\leq \mathbb{E}_{(x,y) \sim \pi}[\|y - x\| I_A] \\
&\leq \mathbb{E}_{(x,y) \sim \pi}[d(x, y)].
\end{aligned}
$$

$$
\begin{aligned}
& W_{LL}(\mathcal{P}_r, \mathcal{P}_g) \leq \mathbb{E}_{(x,y) \sim \pi}[d(x, y)], \forall \pi \in \Pi(\mathcal{P}_r, \mathcal{P}_g) \\
& \Rightarrow W_{LL}(\mathcal{P}_r, \mathcal{P}_g) \leq \inf\nolimits_{\pi \in \Pi(\mathcal{P}_r, \mathcal{P}_g)} \ \mathbb{E}_{(x,y) \sim \pi}\,[d(x, y)] = W_1(\mathcal{P}_r, \mathcal{P}_g).
\end{aligned}
$$

(iii) Combining (i) and (ii), we have $W_{KR}(\mathcal{P}_r, \mathcal{P}_g) \leq W_{LL}(\mathcal{P}_r, \mathcal{P}_g) \leq W_1(\mathcal{P}_r, \mathcal{P}_g)$.

Given $I(\mathcal{P}_r, \mathcal{P}_g) = W_1(\mathcal{P}_r, \mathcal{P}_g)$, we have $I(\mathcal{P}_r, \mathcal{P}_g) = W_{LL}(\mathcal{P}_r, \mathcal{P}_g) = W_1(\mathcal{P}_r, \mathcal{P}_g)$. □

B. The Practical Behaviors of Gradient Uninformativeness

To study the practical behaviors of gradient uninformativeness, we conducted a set of experiments with various hyper-parameter settings. We use the Least-Squares GAN in this experiments as an representative of traditional GANs. The value surface and the gradient of generated samples under various situations are plotted as follows.

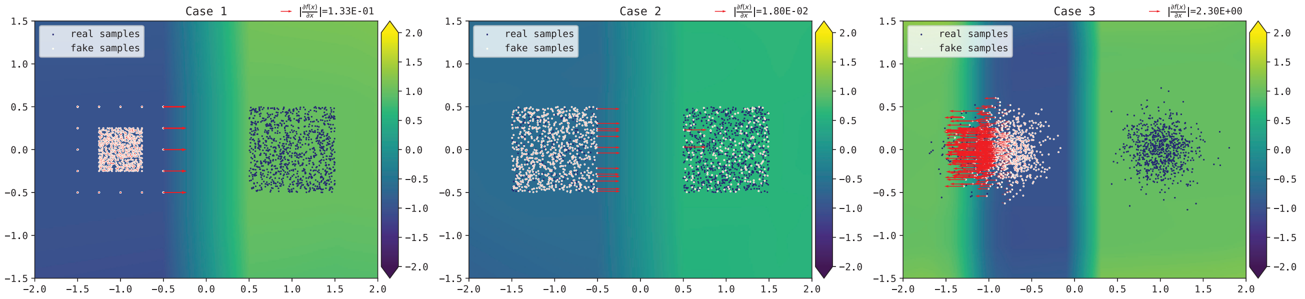


Figure 7: ADAM with lr=1e-2, beta1=0.0, beta2=0.9. MLP with RELU activations, #hidden units=1024, #layers=1.

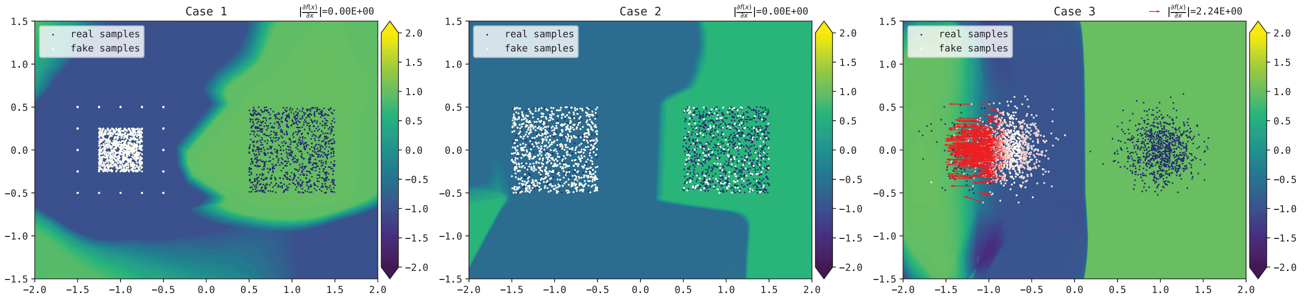


Figure 8: ADAM with lr=1e-2, beta1=0.0, beta2=0.9. MLP with RELU activations, #hidden units=1024, #layers=4.

These experiments shown that the practical f highly depend on the hyper-parameter setting. Given limited capacity, the neural network try to learn the best f. When the neural network is capable of learning approximately the optimal f^*, how the actual f approaches f^* and how the points whose gradients are theoretically undefined behave highly depends the optimization details and the characteristics of the network.

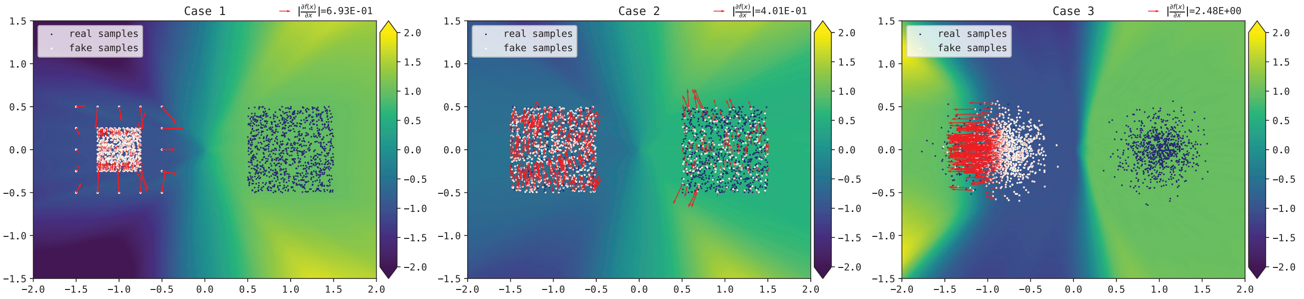


Figure 9: ADAM with lr=1e-5, beta1=0.0, beta2=0.9. MLP with RELU activations, #hidden units=1024, #layers=4.

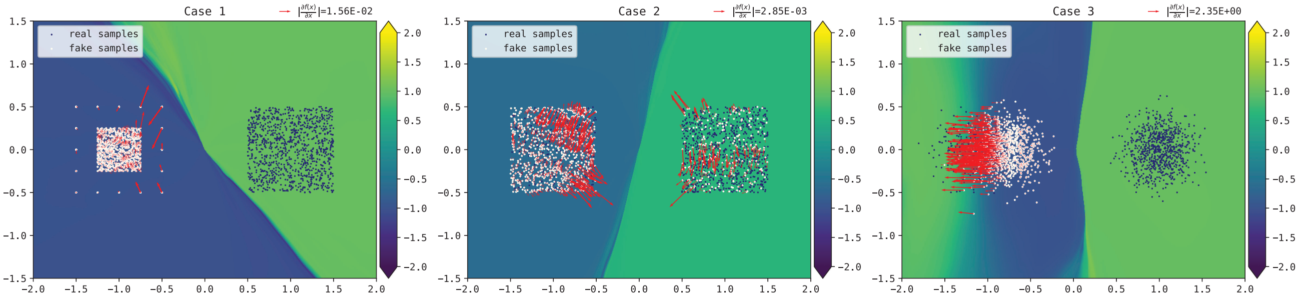


Figure 10: SGD with lr=1e-3. MLP with SELU activations, #hidden units=128, #layers=64.

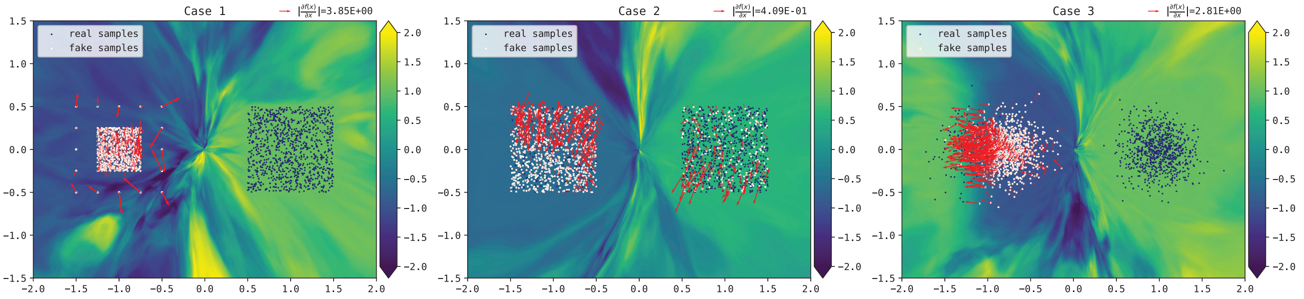


Figure 11: SGD with lr=1e-4. MLP with SELU activations, #hidden units=128, #layers=64.

C. On the Implementation of Lipschitz continuity for GANs

Typical techniques for enforcing k-Lipschitz includes: spectral normalization (Miyato et al., 2018), gradient penalty (Gulrajani et al., 2017), and Lipschitz penalty (Petzka et al., 2017). Before moving into the detailed discussion of these methods, we would like to provide several important notes in the first place.

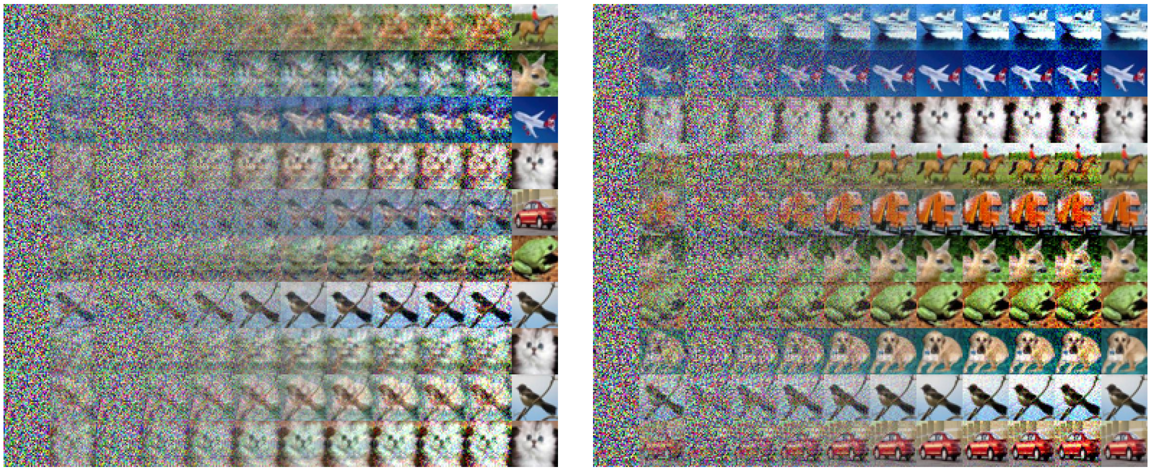
Firstly, enforcing k-Lipschitz in the blending-region of $\mathcal{P}_r$ and $\mathcal{P}_g$ is actually sufficient.

Define $B(\mathcal{S}_r, \mathcal{S}_g) = \{\hat{x} = x \cdot t + y \cdot (1-t) \mid x \in \mathcal{S}_r \text{ and } y \in \mathcal{S}_g \text{ and } t \in [0,1]\}$. It is clear that f is 1-Lipschitz in $B(\mathcal{S}_r, \mathcal{S}_g)$ implies $f(x) - f(y) \leq d(x,y), \forall x \in \mathcal{S}_r, \forall y \in \mathcal{S}_g$. Thus, it is a sufficient constraint for Wasserstein distance in Eq. (4). In fact, $f(x)$ is k-Lipschitz in $B(\mathcal{P}_r, \mathcal{P}_g)$ is also a sufficient condition for all properties described in Lipschitz GANs.

Secondly, enforcing k-Lipschitz with regularization would provide a dynamic Lipschitz constant k.

Lemma 15. *With Wasserstein GAN objective, we have* $\min_{f \in \mathcal{F}_{k\text{-Lip}}} J_D(f) = k \cdot \min_{f \in \mathcal{F}_{1\text{-Lip}}} J_D(f)$.

Assuming we can directly control the Lipschitz constant $k(f)$ of f, the total loss of the discriminator becomes $J(k) \triangleq \min_{f \in \mathcal{F}_{k\text{-Lip}}} J_D(f) + \lambda \cdot (k - k_0)^2$. With Lemma 15, let $\alpha = -\min_{f \in \mathcal{F}_{1\text{-Lip}}} J_D(f)$, then $J(k) = -k \cdot \alpha + \lambda \cdot (k - k_0)^2$, and $J(k)$ achieves its minimum when $k = \frac{\alpha}{2\lambda} + k_0$. When α goes to zero, i.e., $\mathcal{P}_g$ converges to $\mathcal{P}_r$, the optimal k decreases. And when $\mathcal{P}_r = \mathcal{P}_g$, we have $\alpha = 0$ and the optimal $k = k_0$. The similar analysis applies to Lipschitz GANs.



(a) Gradient Penalty

(b) Maximum Gradient Penalty

Figure 12: Comparison between gradient penalty and maximum gradient penalty, with $\mathcal{P}_r$ and $\mathcal{P}_g$ consist of ten real and noise images, respectively. The leftmost in each row is a $x \in \mathcal{S}_g$ and the second is its gradient $\nabla_x f^*(x)$. The interiors are $x + \epsilon \cdot \nabla_x f^*(x)$ with increasing ϵ, which will pass through a real sample, and the rightmost is the nearest $y \in \mathcal{S}_r$.

C.1. Existing Methods

For practical methods, though spectral normalization (Miyato et al., 2018) recently demonstrates their excellent results in training GANs, spectral normalization is an absolute constraint for Lipschitz over the entire space, i.e., constricting the maximum gradient of the entire space, which is unnecessary. On the other side, we also notice both penalty methods proposed in (Gulrajani et al., 2017) and (Petzka et al., 2017) are not exact implementation of the Lipschitz continuity condition, because it does not directly penalty the maximum gradient, but penalties all gradients towards the given target Lipschitz constant or penalties all these greater than one towards the given target.

We also empirically found that the existing methods including spectral normalization (Miyato et al., 2018), gradient penalty (Gulrajani et al., 2017), and Lipschitz penalty (Petzka et al., 2017) all fail to converge to the optimal $f^*(x)$ in some of our synthetic experiments.

C.2. The New Method

Note that this practical method of imposing Lipschitz continuity is not the key contribution of this work. We leave the more rigorous study on this topic as our further work. We introduce it for the necessity for understanding our paper and reproducing of experiments.

Combining the idea of spectral normalization and gradient penalty, we developed a new way of implementing the regularization of Lipschitz continuity in our experiments. Spectral normalization is actually constraining the maximum gradient over the entire space. And as we argued previously, enforcing Lipschitz continuity in the blending region is sufficient. Therefore, we propose to restricting the maximum gradient over the blending region:

$$J_{\text{maxgp}} = \lambda \max_{x \sim B(\mathcal{S}_r, \mathcal{S}_g)} [\big\|\nabla_x f(x)\big\|^2] \tag{27}$$

In practice, we sample x from $B(\mathcal{S}_r, \mathcal{S}_g)$ as in (Gulrajani et al., 2017; Petzka et al., 2017) using training batches of real and fake samples.

We compare the practical result of (centralized) gradient penalty $\mathbb{E}_{x \sim B}[\big\|\nabla_x f(x)\big\|^2]$ and the proposed maximum gradient penalty in Figure 12. Before switching to maximum gradient penalty, we struggled for a long time and cannot achieve a high quality result as shown in Figure 12b. The other forms of gradient penalty (Gulrajani et al., 2017; Petzka et al., 2017) perform similar as $\mathbb{E}_{x \sim B}[\big\|\nabla_x f(x)\big\|^2]$.

To improve the stability and reduce the bias introduced via batch sampling, one can further keep track x with the maximum $\big\|\nabla_x f(x)\big\|$. A practical and light weight method is to maintain a list S_{max} that has the currently highest (top-k) $\big\|\nabla_x f(x)\big\|_2$

(initialized with random samples), use the S_{max} as part of the batch that estimates J_{maxgp}, and update the S_{max} after each batch updating of the discriminator. According to our experiments, it is usually does not improve the training significantly.

D. Extended Discussions and More Details

D.1. Various ϕ and φ That Satisfies Eq. (11)

For Lipschitz GANs, ϕ and φ are required to satisfy Eq. (11). Eq. (11) is actually quite general and there exists many other instances, e.g., $\phi(x) = \varphi(-x) = x$, $\phi(x) = \varphi(-x) = -\log(\sigma(-x))$, $\phi(x) = \varphi(-x) = x + \sqrt{x^2 + \alpha}$ with $\alpha > 0$, $\phi(x) = \varphi(-x) = \exp(x)$, etc. We plot these instances of ϕ and φ in Figure 13.

To devise a loss satisfies Eq. (11), it is practical to let ϕ be an increasing function with non-decreasing derivative and set $\phi(x) = \varphi(-x)$. Note that rescaling and offsetting along the axes are trivial operation to found more ϕ and φ within a function class, and linear combination of two or more ϕ or φ from different function classes also keep satisfying Eq. (11).

D.2. Experiment Details

In our experiments with real datas (CIFAR-10, Tiny Imagenet and Oxford 102), we follow the network architecture and hyper-parameters in (Gulrajani et al., 2017). The network architectures are detailed in Table 3. We use Adam optimizer with beta1=0.0, beta2=0.9, and the learning rate is 0.0002 which linear decays to zero in 200, 000 iterations. We use 5 discriminator updates per generator update. We use MaxGP for all our experiments of LGANs and search the best penalty weight λ in $[0.01, 0.1, 1.0, 10.0]$. Please check more details in our codes. For all experiments in Table 2, we only change ϕ and φ and the dataset, and all other components are fixed.

We plot the IS training curve of LGANs in Figure 14 and 15. We provide the visual results of LGANs in Figure 16, Figure 17 for CIFAR-10 and Tiny Imagenet, respectively. As an extra experiment, we also provide the visual results of LGANs on Oxford 102 in Figure 18.

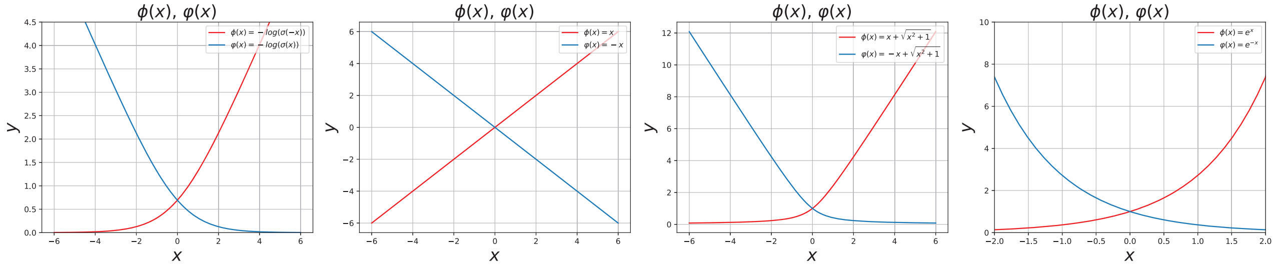


Figure 13: Various ϕ and φ that satisfies Eq. (11).

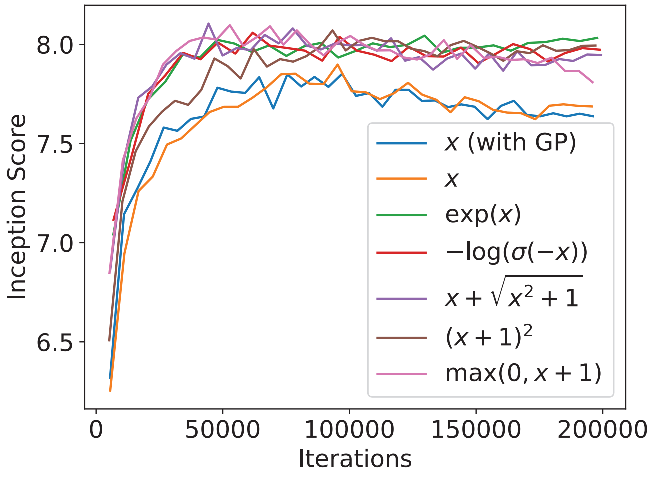


Figure 14: IS training curves on CIFAR-10.

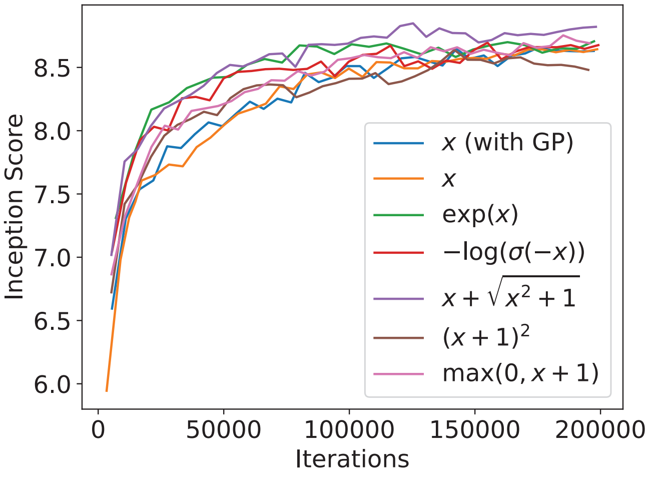
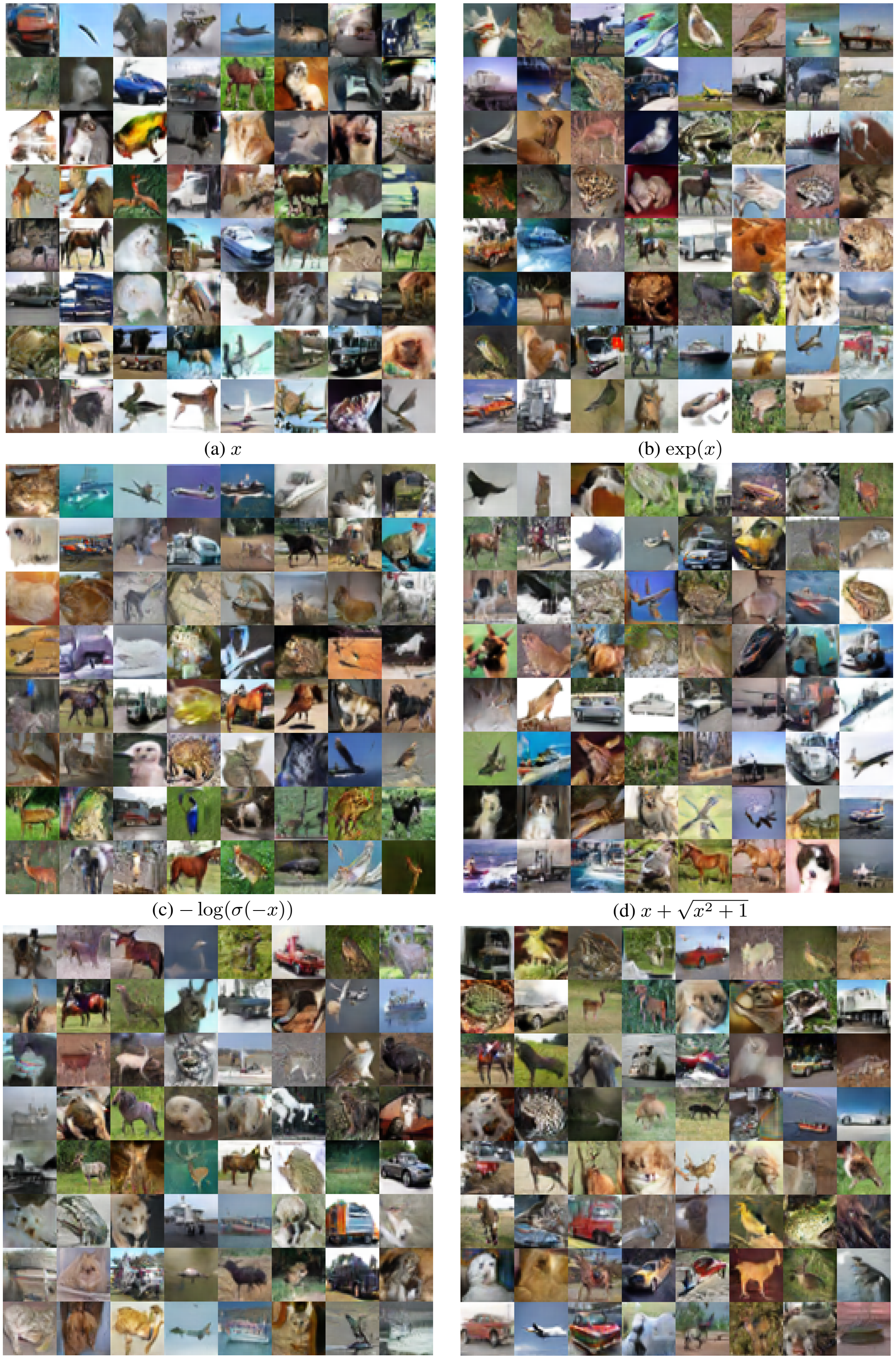


Figure 15: IS training curves on Tiny ImageNet.



(a) x

(b) $\exp(x)$

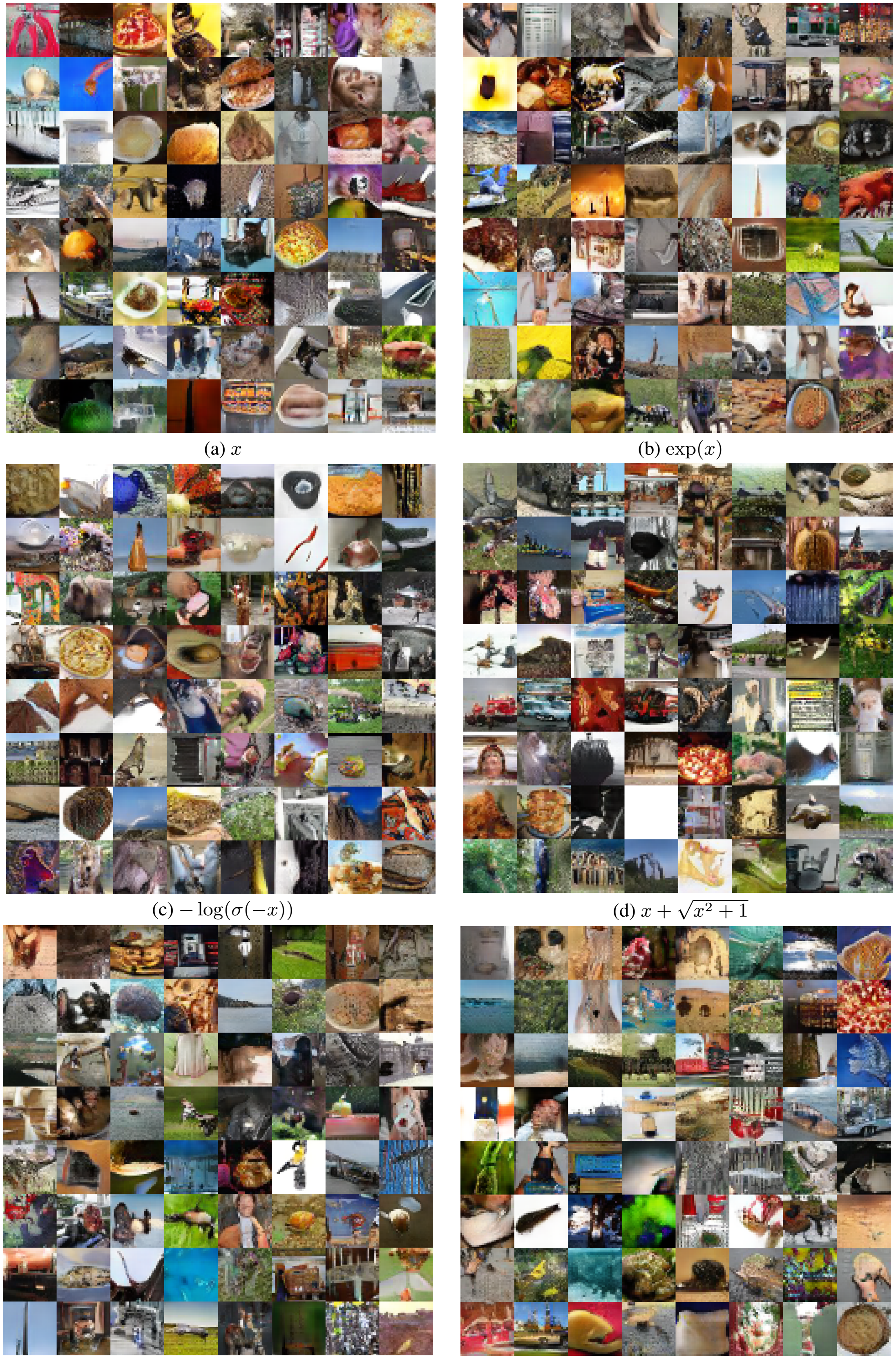
(c) $-\log(\sigma(-x))$

(d) $x + \sqrt{x^2 + 1}$

(e) $(x + 1.0)^2$

(f) $\max(0, x + 1.0)$

Figure 16: Random samples of LGANs with different loss metrics on CIFAR-10.



(a) x

(b) $\exp(x)$

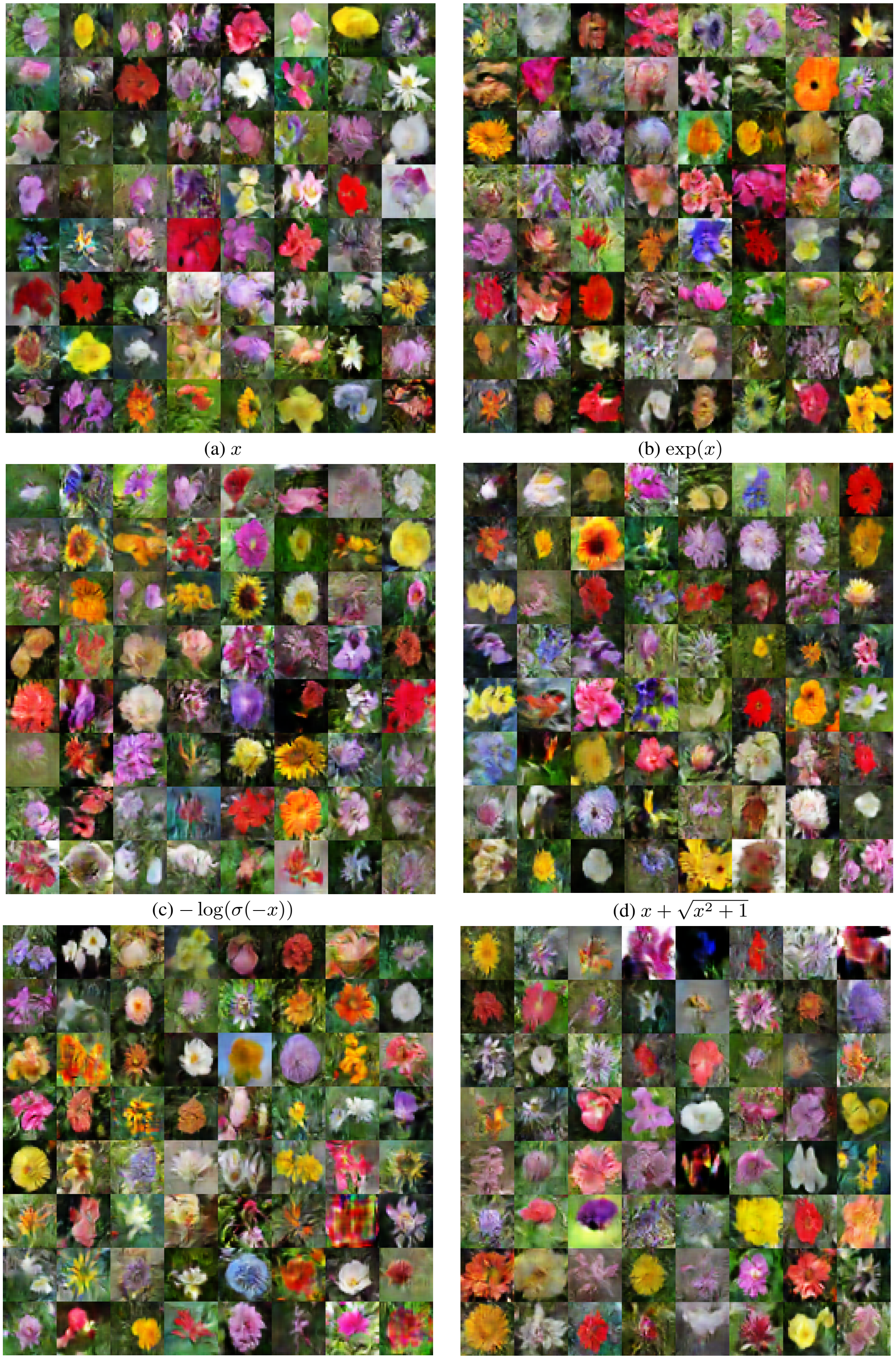
(c) $-\log(\sigma(-x))$

(d) $x + \sqrt{x^2 + 1}$

(e) $(x + 1.0)^2$

(f) $\max(0, x + 1.0)$

Figure 17: Random samples of LGANs with different loss metrics on Tiny Imagenet.



(a) x

(b) $\exp(x)$

(c) $-\log(\sigma(-x))$

(d) $x + \sqrt{x^2 + 1}$

(e) $(x + 1.0)^2$

(f) $\max(0, x + 1.0)$

Figure 18: Random samples of LGANs with different loss metrics on Oxford 102.

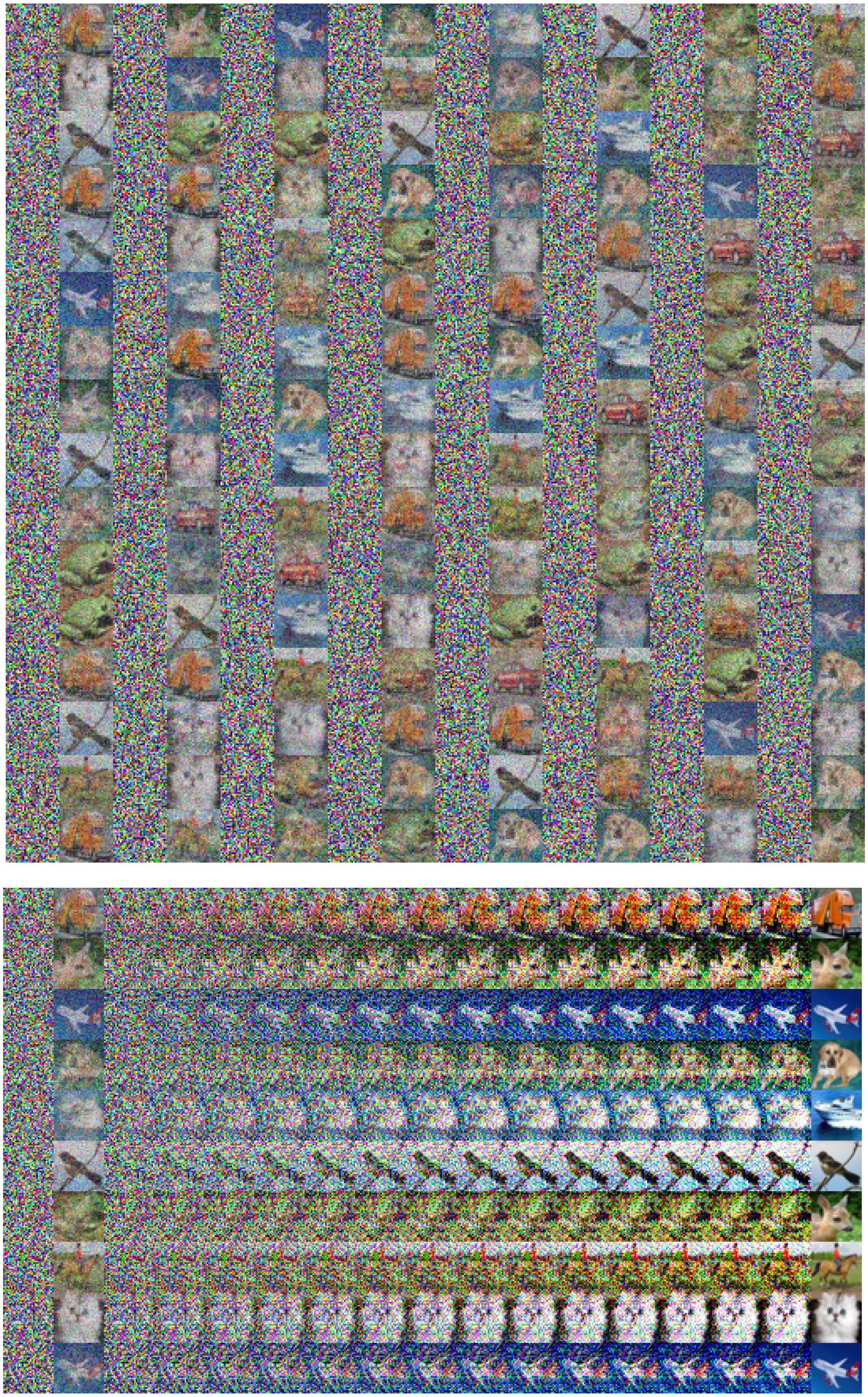


Figure 19: The gradient of LGANs with real world data, where $\mathcal{P}_r$ consists of ten images and $\mathcal{P}_g$ is Gaussian noise. Up: Each odd column are $x \in \mathcal{S}_g$ and the nearby column are their gradient $\nabla_x f^*(x)$. Down: the leftmost in each row is $x \in \mathcal{S}_g$, the second are their gradients $\nabla_x f^*(x)$, the interiors are $x + \epsilon \cdot \nabla_x f^*(x)$ with increasing ϵ, and the rightmost is the nearest $y \in \mathcal{S}_r$.

Generator:

Operation	Kernel	Resample	Output Dims
Noise	N/A	N/A	128
Linear	N/A	N/A	128×4×4
Residual block	3×3	UP	128×8×8
Residual block	3×3	UP	128×16×16
Residual block	3×3	UP	128×32×32
Conv & Tanh	3×3	N/A	3×32×32

Discriminator:

Operation	Kernel	Resample	Output Dims
Residual Block	3×3×2	Down	128×16×16
Residual Block	3×3×2	Down	128×8×8
Residual Block	3×3×2	N/A	128×8×8
Residual Block	3×3×2	N/A	128×8×8
ReLU,mean pool	N/A	N/A	128
Linear	N/A	N/A	1

Table 3: The network architectures.