A. Proof of Things

Proof of Theorem 1. The proof of this theorem is split into smaller lemmas that are proven individually.

- That τ_P is a strict adversarial divergence which is equivalent to τ_W is proven in Lemma 4, thus showing that τ_P fulfills Requirement 1.
- τ_P fulfills Requirement 2 by design.
- The existence of an optimal critic in $\mathrm{OC}_{\tau_P}(\mathbb{P},\mathbb{Q})$ follows directly from Lemma 3.
- That there exists a critic $f^* \in \mathrm{OC}_{\tau_P}(\mathbb{P},\mathbb{Q})$ that fulfills Eq. 5 is because Lemma 3 ensures that a continuous differentiable f^* exists in $\mathrm{OC}_{\tau_P}(\mathbb{P},\mathbb{Q})$ which fulfills Eq. 9. Because Eq. 9 holds for $f^* \in C(X)$, the same reasoning as the end of the proof of Lemma 7 can be used to show Requirement 4

□

We prepare by showing a few basic lemmas used in the remaining proofs

Lemma 1 (concavity of $\tau_P(\mathbb{P}\|\mathbb{Q};\cdot)$)**.** *The mapping $C^1(X) \to \mathbb{R}$, $f \mapsto \tau_P(\mathbb{P}\|\mathbb{Q}; f)$ is concave.*

Proof. The concavity of $f \mapsto \mathbb{E}_{x\sim\mathbb{P}}[f(x)] - \mathbb{E}_{x'\sim\mathbb{Q}}[f(x')]$ is trivial. Now consider $\gamma \in (0,1)$, then

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathbb{P},x'\sim\mathbb{Q}}\left[\frac{(\gamma(f(x)-f(x'))+(1-\gamma)(\hat{f}(x)-\hat{f}(x')))^2}{\|x-x'\|}\right]\\
\leq&\mathbb{E}_{x\sim\mathbb{P},x'\sim\mathbb{Q}}\left[\frac{\gamma(f(x)-f(x'))^2+(1-\gamma)(\hat{f}(x)-\hat{f}(x'))^2}{\|x-x'\|}\right]\\
=&\gamma\mathbb{E}_{x\sim\mathbb{P},x'\sim\mathbb{Q}}\left[\frac{(f(x)-f(x'))^2}{\|x-x'\|}\right]+(1-\gamma)\mathbb{E}_{x\sim\mathbb{P},x'\sim\mathbb{Q}}\left[\frac{(\hat{f}(x)-\hat{f}(x'))^2}{\|x-x'\|}\right],
\end{aligned}$$

thus showing concavity of $\tau_P(\mathbb{P}\|\mathbb{Q};\cdot)$. □

Lemma 2 (necessary and sufficient condition for maximum)**.** *Assume $\mathbb{P},\mathbb{Q} \in \mathcal{P}(X)$ fulfill assumptions 1 and 2. Then for any $f \in \mathrm{OC}_{\tau_P}(\mathbb{P},\mathbb{Q})$ it must hold that*

$$P_{x'\sim\mathbb{Q}}\left(\mathbb{E}_{x\sim\mathbb{P}}\left[\frac{f(x)-f(x')}{\|x-x'\|}\right]=\frac{1}{2\lambda}\right)=1 \tag{7}$$

and

$$P_{x\sim\mathbb{P}}\left(\mathbb{E}_{x'\sim\mathbb{Q}}\left[\frac{f(x)-f(x')}{\|x-x'\|}\right]=\frac{1}{2\lambda}\right)=1. \tag{8}$$

Further, if $f \in C^1(X)$ and fulfills Eq. 7 and 8, then $f \in \mathrm{OC}_{\tau_P}(\mathbb{P},\mathbb{Q})$

Proof. Since in Lemma 1 it was shown that the the mapping $f \mapsto \tau_P(\mathbb{P}\|\mathbb{Q}, f)$ is concave, $f \in \mathrm{OC}_\tau(\mathbb{P},\mathbb{Q})$ if and only if $f \in C^1(X)$ and f is a local maximum of $\tau_P(\mathbb{P}\|\mathbb{Q};\cdot)$. This is equivalent to saying that all $u_1, u_2 \in C^1(X)$ with $\mathrm{supp}(u_1) \cap \mathrm{supp}(\mathbb{Q}) = \emptyset$ and $\mathrm{supp}(u_2) \cap \mathrm{supp}(\mathbb{P}) = \emptyset$ it holds

$$\nabla_{(\varepsilon,\rho)}\left[\mathbb{E}_\mathbb{P}[f+\varepsilon u_1]-\mathbb{E}_\mathbb{Q}[f+\rho u_2]-\lambda\mathbb{E}_{x\sim\mathbb{P},x'\sim\mathbb{Q}}\left[\frac{((f+\varepsilon u_1)(x)-(f+\rho u_2)(x'))^2}{\|x-x'\|}\right]\right]\Bigg|_{\varepsilon=0,\rho=0}=0$$

which holds if and only if

$$\mathbb{E}_{x\sim\mathbb{P}}\left[u_1(x)\left(1-2\lambda\mathbb{E}_{x'\sim\mathbb{Q}}\left[\frac{(f(x)-f(x'))}{\|x-x'\|}\right]\right)\right]=0$$

and

$$\mathbb{E}_{x'\sim\mathbb{Q}}\left[u_2(x')\left(1-2\lambda\mathbb{E}_{x\sim\mathbb{P}}\left[\frac{(f(x)-f(x'))}{\|x-x'\|}\right]\right)\right]=0$$

proving that Eq. 7 and 8 are necessary and sufficient. □

Lemma 3. *Let $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(X)$ be probability measures fulfilling Assumptions 1 and 2. Define an open subset of X, $\Omega \subseteq X$, such that $\operatorname{supp}(\mathbb{Q}) \subseteq \Omega$ and $\inf_{x \in \operatorname{supp}(\mathbb{P}), x' \in \Omega} \|x - x'\| > 0$. Then there exists a $f \in \mathcal{F} = C^1(X)$ such that*

$$\forall x' \in \Omega: \quad \mathbb{E}_{x \sim \mathbb{P}}\left[\frac{f(x) - f(x')}{\|x - x'\|}\right] = \frac{1}{2\lambda} \tag{9}$$

and

$$\forall x \in \operatorname{supp}(\mathbb{P}): \quad \mathbb{E}_{x' \sim \mathbb{Q}}\left[\frac{f(x) - f(x')}{\|x - x'\|}\right] = \frac{1}{2\lambda} \tag{10}$$

and $\tau_P(\mathbb{P}\|\mathbb{Q}; f) = \tau_P(\mathbb{P}\|\mathbb{Q})$.

Proof. Since $\tau(\mathbb{P}\|\mathbb{Q}; f) = \tau(\mathbb{P}\|\mathbb{Q}; f + c)$ for any $c \in \mathbb{R}$ and is only affected by values of f on $\operatorname{supp}(\mathbb{P}) \cup \Omega$ we first start by considering

$$\mathcal{F} = \left\{ f \in C^1(\operatorname{supp}(\mathbb{P}) \cup \Omega) \mid \mathbb{E}_{x \sim \mathbb{P}, x' \sim \mathbb{Q}}\left[\frac{f(x')}{\|x - x'\|}\right] = 0 \right\}$$

Observe that Eq. 9 holds if

$$x' \in \Omega: \quad f(x') = \frac{\mathbb{E}_{x \sim \mathbb{P}}[\frac{f(x)}{\|x - x'\|}] - \frac{1}{2\lambda}}{\mathbb{E}_{x \sim \mathbb{P}}[\frac{1}{\|x - x'\|}]}$$

and similarly for Eq. 10

$$\forall x \in \operatorname{supp}(\mathbb{P}): \quad f(x) = \frac{\mathbb{E}_{x' \sim \mathbb{Q}}[\frac{f(x')}{\|x - x'\|}] + \frac{1}{2\lambda}}{\mathbb{E}_{x' \sim \mathbb{Q}}[\frac{1}{\|x - x'\|}]}.$$

Now it's clear that if the mapping $T : \mathcal{F} \to \mathcal{F}$ defined by

$$T(f)(x) := \begin{cases} \frac{\mathbb{E}_{x' \sim \mathbb{Q}}[\frac{f(x')}{\|x - x'\|}] + \frac{1}{2\lambda}}{\mathbb{E}_{x' \sim \mathbb{Q}}[\frac{1}{\|x - x'\|}]} & x \in \operatorname{supp}(\mathbb{P}) \\ \frac{\mathbb{E}_{x' \sim \mathbb{P}}[\frac{f(x')}{\|x - x'\|}] - \frac{1}{2\lambda}}{\mathbb{E}_{x' \sim \mathbb{P}}[\frac{1}{\|x - x'\|}]} & x \in \Omega \end{cases} \tag{11}$$

admit a fix point $f^* \in \mathcal{F}$, i.e. $T(f^*) = f^*$, then f^* is a solution to Eq. 9 and 10, and with that a solution to Eq. 7 and 8 and $\tau_P(\mathbb{P}\|\mathbb{Q}; f^*) = \tau_P(\mathbb{P}\|\mathbb{Q})$.

Define the mapping $S : \mathcal{F} \to \mathcal{F}$ by

$$S(f)(x) = \frac{f(x)}{2\lambda \mathbb{E}_{\tilde{x} \sim \mathbb{P}, x' \sim \mathbb{Q}}\left[\frac{f(\tilde{x}) - f(x')}{\|\tilde{x} - x'\|}\right]}.$$

Then

$$\mathbb{E}_{\tilde{x} \sim \mathbb{P}, x' \sim \mathbb{Q}}\left[\frac{S(f)(\tilde{x}) - S(f)(x')}{\|\tilde{x} - x'\|}\right] = \frac{1}{2\lambda} \tag{12}$$

and

$$S(S(f))(x) = \frac{S(f)(x)}{2\lambda \mathbb{E}_{\tilde{x} \sim \mathbb{P}, x' \sim \mathbb{Q}}\left[\frac{S(f)(\tilde{x}) - S(f)(x')}{\|\tilde{x} - x'\|}\right]} = \frac{S(f)(x)}{2\lambda \frac{1}{2\lambda}} = S(f)(x)$$

making S a projection. By the same reasoning, if $\mathbb{E}_{\tilde{x} \sim \mathbb{P}, x' \sim \mathbb{Q}}\left[\frac{f(\tilde{x}) - f(x')}{\|\tilde{x} - x'\|}\right] = \frac{1}{2\lambda}$ then f is a fix-point of S, i.e. $S(f) = f$. Assume f is such a function, then by definition of T in Eq. 11

$$\begin{aligned}
\mathbb{E}_{\tilde{x} \sim \mathbb{P}, x' \sim \mathbb{Q}}\left[\frac{T(f)(\tilde{x}) - T(f)(x')}{\|\tilde{x} - x'\|}\right] &= \mathbb{E}_{\tilde{x} \sim \mathbb{P}}\left[\mathbb{E}_{x' \sim \mathbb{Q}}\left[\frac{T(f)(\tilde{x})}{\|\tilde{x} - x'\|}\right]\right] - \mathbb{E}_{x' \sim \mathbb{Q}}\left[\mathbb{E}_{\tilde{x} \sim \mathbb{P}}\left[\frac{T(f)(x')}{\|\tilde{x} - x'\|}\right]\right] \\
&= \mathbb{E}_{\tilde{x} \sim \mathbb{P}}\left[\mathbb{E}_{x' \sim \mathbb{Q}}\left[\frac{f(x')}{\|\tilde{x} - x'\|}\right] + \frac{1}{2\lambda}\right] - \mathbb{E}_{x' \sim \mathbb{Q}}\left[\mathbb{E}_{\tilde{x} \sim \mathbb{P}}\left[\frac{f(\tilde{x})}{\|\tilde{x} - x'\|}\right] - \frac{1}{2\lambda}\right] \\
&= -\mathbb{E}_{\tilde{x} \sim \mathbb{P}, x' \sim \mathbb{Q}}\left[\frac{f(\tilde{x}) - f(x')}{\|\tilde{x} - x'\|}\right] + 2\frac{1}{2\lambda} \\
&= \frac{1}{2\lambda}.
\end{aligned}$$

Therefore, $S(T(S(f))) = T(S(f))$. We can define $S(\mathcal{F}) = \{S(f) \mid f \in \mathcal{F}\}$ and see that $T : S(\mathcal{F}) \to S(\mathcal{F})$. Further, since $S(\cdot)$ only multiplies with a scalar, $S(F) \subseteq \mathcal{F}$.

Let $f_1, f_2 \in S(\mathcal{F})$. From Eq. 12 we get

$$\mathbb{E}_{x\sim\mathbb{P},x'\sim\mathbb{Q}}\left[\frac{f_1(x') - f_2(x')}{\|x - x'\|}\right] = \mathbb{E}_{x\sim\mathbb{P},x'\sim\mathbb{Q}}\left[\frac{f_1(x) - f_2(x)}{\|x - x'\|}\right].$$

Now since for every $f \in \mathcal{F}$ it holds by design that $\mathbb{E}_{x\sim\mathbb{P},x'\sim\mathbb{Q}}\left[\frac{f(x')}{\|x-x'\|}\right] = 0$ and since $S(\mathcal{F}) \subseteq \mathcal{F}$ we see that $f_1, f_2 \in S(\mathcal{F})$ that

$$\mathbb{E}_{x\sim\mathbb{P},x'\sim\mathbb{Q}}\left[\frac{f_1(x') - f_2(x')}{\|x - x'\|}\right] = \mathbb{E}_{x\sim\mathbb{P},x'\sim\mathbb{Q}}\left[\frac{f_1(x) - f_2(x)}{\|x - x'\|}\right] = 0$$

Using this with the continuity of f_1, f_2, there must exist $x_1 \in \mathrm{supp}(\mathbb{P})$ with

$$\mathbb{E}_{x'\sim\mathbb{Q}}\left[\frac{f_1(x') - f_2(x')}{\|x_1 - x'\|}\right] = 0$$

With this (and compactness of our domain), $\mathbb{Q}$ must have mass in both positive and negative regions of $f_1 - f_2$ and exists a constant $p < 1$ such that for all $f_1, f_2 \in S(\mathcal{F})$ it holds

$$\sup_{x\in\mathrm{supp}(\mathbb{P})}\left|\mathbb{E}_{x'\sim\mathbb{Q}}\left[\frac{f_1(x') - f_2(x')}{\|x - x'\|}\right]\right| \le p \sup_{x\in\mathrm{supp}(\mathbb{P})} \mathbb{E}_{x'\sim\mathbb{Q}}\left[\frac{1}{\|x - x'\|}\right] \sup_{x'\in\Omega} |f_1(x') - f_2(x')|. \tag{13}$$

To show the existence of a fix-point for T in the Banach Space $(\mathcal{F}, \|\cdot\|_\infty)$ we use the Banach fixed-point theorem to show that T has a fixed point in the metric space $(S(\mathcal{F}), \|\cdot\|_\infty)$ (remember that $T : S(\mathcal{F}) \to S(\mathcal{F})$ and $S(\mathcal{F}) \subseteq \mathcal{F}$). If $f_1, f_2 \in S(\mathcal{F})$ then

$$\begin{aligned}\sup_{x\in\mathrm{supp}(\mathbb{P})} |T(f_1)(x) - T(f_2)(x)| &= \sup_{x\in\mathrm{supp}(\mathbb{P})}\left|\frac{\mathbb{E}_{x'\sim\mathbb{Q}}\left[\frac{f_1(x')-f_2(x')}{\|x-x'\|}\right]}{\mathbb{E}_{x'\sim\mathbb{Q}}\left[\frac{1}{\|x-x'\|}\right]}\right| \\ &\le p \sup_{x'\in\mathrm{supp}(\mathbb{Q})} |f_1(x') - f_2(x')| \quad \text{using Eq. 13}\end{aligned}$$

The same trick can be used to find some some $q < 1$ and show

$$\sup_{x'\in\Omega} |T(f_1)(x') - T(f_2)(x')| \le q \sup_{x\in\mathrm{supp}(\mathbb{P})} |f_1(x) - f_2(x)|$$

thereby showing

$$\|T(f_1) - T(f_2)\|_\infty < \max(p, q)\|f_1 - f_2\|_\infty$$

The Banach fix-point theorem then delivers the existence of a fix-point $f^* \in S(\mathcal{F})$ for T.

Finally, we can use the Tietze extension theorem to extend f^* to all of X, thus finding a fix point for T in $C^1(X)$ and proving the lemma. □

Lemma 4. *τ_P is a strict adversarial divergence and τ_P and τ_W are equivalent.*

Proof. Let $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(X)$ be two probability measures fulfilling Assumptions 1 and 2 with $\mathbb{P} \neq \mathbb{Q}$. It's shown in (Sriperumbudur et al., 2010) that $\mu = \tau_W(\mathbb{P}, \mathbb{Q}) > 0$, meaning there exists a function $f \in C(X)$, $\|f\|_L \le 1$ such that

$$\mathbb{E}_{\mathbb{P}}[f] - \mathbb{E}_{\mathbb{Q}}[f] = \mu > 0.$$

The Stone–Weierstrass theorem tells us that there exists a $f' \in C_\infty(X)$ such that $\|f - f'\|_\infty \le \frac{\mu}{4}$ and thus $\mathbb{E}_{\mathbb{P}}[f'] - \mathbb{E}_{\mathbb{Q}}[f'] \ge \frac{\mu}{2}$. Now consider the function $\varepsilon f'$ with $\varepsilon > 0$, it's clear that

$$\tau_P(\mathbb{P}\|\mathbb{Q}) \ge \tau_P(\mathbb{P}\|\mathbb{Q}; \varepsilon f') = \varepsilon(\underbrace{\mathbb{E}_{\mathbb{P}}[f'] - \mathbb{E}_{\mathbb{Q}}[f']}_{\ge\frac{\mu}{2}}) - \varepsilon^2\lambda\mathbb{E}_{x\sim\mathbb{P},x'\sim\mathbb{Q}}[\frac{(f'(x) - f'(x'))^2}{\|x - x'\|}]$$

and so for a sufficiently small $\varepsilon > 0$ we'll get $\tau_P(\mathbb{P}\|\mathbb{Q}; \varepsilon f') > 0$ meaning $\tau_P(\mathbb{P}\|\mathbb{Q}) > 0$ and τ_P is a strict adversarial divergence.

To show equivalence, we note that

$$\tau_P(\mathbb{P}\|\mathbb{Q}) \leq \sup_{m \in C(X^2)} \mathbb{E}_{x\sim\mathbb{P}, x'\sim\mathbb{Q}} \left[m(x,x') \left(1 - \lambda \frac{m(x,x')}{\|x-x'\|} \right) \right]$$

therefore for any optimum it must hold $m(x,x') \leq \frac{\|x-x'\|}{2\lambda}$, and thus (similar to Lemma 2) any optimal solution will be Lipschitz continuous with a the Lipschitz constant independent of $\mathbb{P}, \mathbb{Q}$. Thus $\tau_W(\mathbb{P}\|\mathbb{Q}) \geq \gamma \tau_P(\mathbb{P}\|\mathbb{Q})$ for $\gamma > 0$, from which we directly get equivalence. □

Proof of Theorem 2. We start by applying Lemma 5 giving us

- $\mathrm{OC}_{\tau_F}(\mathbb{P}, \mathbb{Q}'_{\theta_0}) \neq \emptyset$.
- For any $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(X)$ fulfilling Assumptions 1 and 2, it holds that $\tau_F(\mathbb{P}\|\mathbb{Q}) = \tau_P(\mathbb{P}\|\mathbb{Q})$, meaning τ_F is like τ_P a strict adversarial divergence which is equivalent to τ_W, showing Requirement 1.
- τ_F fulfills Requirement 2 by design.
- Every $f^* \in \mathrm{OC}_{\tau_F}(\mathbb{P}, \mathbb{Q}'_{\theta_0})$ is in $\mathrm{OC}_{\tau_P}(\mathbb{P}, \mathbb{Q}'_{\theta_0}) \subseteq C^1(X)$, therefore f^* the gradient $\nabla_\theta \mathbb{E}_{\mathbb{Q}_\theta}[f^*]|_{\theta_0}$ exists. Further Lemma 7 shows that the update rule $\nabla_\theta \mathbb{E}_{\mathbb{Q}_\theta}[f^*]|_{\theta_0}$ is unique, thus showing Requirement 3.
- Lemma 7 gives us every $f^* \in \mathrm{OC}_{\tau_F}(\mathbb{P}, \mathbb{Q}'_{\theta_0})$ with the corresponding update rule fulfills Requirement 4, thus proving Theorem 2.

□

Before we can show this theorem, we must prove a few interesting lemmas about τ_F. The following lemma is quite powerful; since $\tau_P(\mathbb{P}\|\mathbb{Q}) = \tau_F(\mathbb{P}\|\mathbb{Q})$ and $\mathrm{OC}_{\tau_F}(\mathbb{P}, \mathbb{Q}) \subseteq \mathrm{OC}_{\tau_P}(\mathbb{P}, \mathbb{Q})$ any property that's proven for τ_P automatically holds for τ_F.

Lemma 5. *If let $X \subseteq \mathbb{R}^n$ and $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(X)$ be probability measures fulfilling Assumptions 1 and 2. Then*

1. *there exists $f^* \in \mathrm{OC}_{\tau_P}(\mathbb{P}, \mathbb{Q})$ so that $\tau_F(\mathbb{P}\|\mathbb{Q}; f^*) = \tau_P(\mathbb{P}\|\mathbb{Q}; f^*)$,*
2. *$\tau_P(\mathbb{P}\|\mathbb{Q}) = \tau_F(\mathbb{P}\|\mathbb{Q})$,*
3. *$\emptyset \neq \mathrm{OC}_{\tau_F}(\mathbb{P}, \mathbb{Q})$,*
4. *$\mathrm{OC}_{\tau_F}(\mathbb{P}, \mathbb{Q}) \subseteq \mathrm{OC}_{\tau_P}(\mathbb{P}, \mathbb{Q})$.*

Clain (4) is especially helpful, now anything that has been proven for all $f^* \in \mathrm{OC}_{\tau_P}(\mathbb{P}, \mathbb{Q})$ automatically holds for all $f^* \in \mathrm{OC}_{\tau_F}(\mathbb{P}, \mathbb{Q})$

Proof. For convenience define

$$G(\mathbb{P}, \mathbb{Q}; f) := \mathbb{E}_{x'\sim\mathbb{Q}} \left[\left(\|\nabla_x f(x)\big|_{x'}\| - \frac{\left\| \mathbb{E}_{\tilde{x}\sim\mathbb{P}}[(\tilde{x} - x') \frac{f(\tilde{x}) - f(x')}{\|x' - \tilde{x}\|^3}] \right\|}{\mathbb{E}_{\tilde{x}\sim\mathbb{P}}[\frac{1}{\|x' - \tilde{x}\|}]} \right)^2 \right]$$

(G is for gradient penalty) and note that

$$\tau_F(\mathbb{P}\|\mathbb{Q}; f) = \tau_P(\mathbb{P}\|\mathbb{Q}; f) - \underbrace{G(\mathbb{P}, \mathbb{Q}; f)}_{\geq 0}$$

Therefore it's clear that $\tau_F(\mathbb{P}\|\mathbb{Q}) \leq \tau_P(\mathbb{P}\|\mathbb{Q})$

Claim (1). Let $\Omega \subseteq X$ be an open set such that $\operatorname{supp}(\mathbb{Q}) \subseteq \Omega$ and $\Omega \cap \operatorname{supp}(\mathbb{P}) = \emptyset$. Then Lemma 3 tells us there is a $f \in \mathrm{OC}_{\tau_P}(\mathbb{P}, \mathbb{Q})$ (and thus $f \in C^1(X)$) such that

$$\forall x' \in \Omega : \quad \mathbb{E}_{\tilde{x} \sim \mathbb{P}} \left[\frac{f(\tilde{x}) - f(x')}{\|\tilde{x} - x'\|} \right] = \frac{1}{2\lambda}$$

and thus, because $\operatorname{supp}(\mathbb{Q}) \subseteq \Omega$ open and $f \in C^1(X)$,

$$\forall x' \in \operatorname{supp}(\mathbb{Q}) : \quad \nabla_x \mathbb{E}_{\tilde{x} \sim \mathbb{P}} \left[\frac{f(\tilde{x}) - f(x)}{\|\tilde{x} - x\|} \right] \Bigg|_{x'} = 0$$

Now taking the gradients with respect to x' gives us

$$\nabla_x \mathbb{E}_{\tilde{x} \sim \mathbb{P}} \left[\frac{f(\tilde{x}) - f(x)}{\|\tilde{x} - x\|} \right] \Bigg|_{x'} = -\nabla_x f(x)|_{x'} \mathbb{E}_{\tilde{x} \sim \mathbb{P}} \left[\frac{1}{\|\tilde{x} - x'\|} \right] + \mathbb{E}_{\tilde{x} \sim \mathbb{P}} \left[(\tilde{x} - x') \frac{f(\tilde{x}) - f(x')}{\|\tilde{x} - x'\|^3} \right] \tag{14}$$

meaning

$$\forall x' \in \operatorname{supp}(\mathbb{Q}) : \quad \nabla_x f(x)|_{x'} = \frac{\mathbb{E}_{\tilde{x} \sim \mathbb{P}} \left[(\tilde{x} - x') \frac{f(\tilde{x}) - f(x')}{\|\tilde{x} - x'\|^3} \right]}{\mathbb{E}_{\tilde{x} \sim \mathbb{P}} \left[\frac{1}{\|\tilde{x} - x'\|} \right]} \tag{15}$$

thus $G(\mathbb{P}, \mathbb{Q}; f) = 0$, showing the claim.

Claims (2) and (3). The claims are a direct result of Claim (1); for every $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(X)$ there exists a

$$f^* \in \mathrm{OC}_{\tau_P}(\mathbb{P}, \mathbb{Q})$$

such that $G(\mathbb{P}, \mathbb{Q}; f^*) = 0$. Therefore

$$\tau_P(\mathbb{P} \| \mathbb{Q}) \geq \tau_F(\mathbb{P} \| \mathbb{Q}) \geq \tau_F(\mathbb{P} \| \mathbb{Q}; f^*) = \tau_P(\mathbb{P} \| \mathbb{Q}; f^*) = \tau_P(\mathbb{P} \| \mathbb{Q})$$

thus showing both $\tau_P(\mathbb{P} \| \mathbb{Q}) = \tau_F(\mathbb{P} \| \mathbb{Q})$ and $f^* \in \mathrm{OC}_{\tau_F}(\mathbb{P} \| \mathbb{Q})$.

Claim (4). This claim is a direct result of claim (2); since $\tau_P(\mathbb{P} \| \mathbb{Q}) = \tau_F(\mathbb{P} \| \mathbb{Q})$, that means that if $f^* \in \mathrm{OC}_{\tau_F}(\mathbb{P} \| \mathbb{Q})$, then

$$\tau_F(\mathbb{P} \| \mathbb{Q}) = \tau_F(\mathbb{P} \| \mathbb{Q}; f^*) = \tau_P(\mathbb{P} \| \mathbb{Q}; f^*) - \underbrace{G(\mathbb{P}, \mathbb{Q}; f)}_{\geq 0} \leq \tau_P(\mathbb{P} \| \mathbb{Q}; f^*) \leq \tau_P(\mathbb{P} \| \mathbb{Q}) = \tau_F(\mathbb{P} \| \mathbb{Q})$$

thus $\tau_P(\mathbb{P} \| \mathbb{Q}; f^*) = \tau_P(\mathbb{P} \| \mathbb{Q})$ and $f^* \in \mathrm{OC}_{\tau_P}(\mathbb{P} \| \mathbb{Q})$. □

Lemma 6. *For every $f^* \in \mathrm{OC}_{\tau_F}(\mathbb{P}, \mathbb{Q}')$ it holds*

$$\forall x' \in \operatorname{supp}(\mathbb{Q}') : \quad \nabla_x \mathbb{E}_{\tilde{x} \sim \mathbb{P}} \left[\frac{f^*(\tilde{x}) - f^*(x)}{\|\tilde{x} - x\|} \right] \Bigg|_{x'} = 0 \tag{16}$$

Proof. Set

$$v = \frac{\mathbb{E}_{\tilde{x} \sim \mathbb{P}}[(\tilde{x} - x') \frac{f^*(\tilde{x}) - f^*(x')}{\|x' - \tilde{x}\|^3}]}{\left\| \mathbb{E}_{\tilde{x} \sim \mathbb{P}}[(\tilde{x} - x') \frac{f^*(\tilde{x}) - f^*(x')}{\|x' - \tilde{x}\|^3}] \right\|}$$

and note that due to construction of $\mathbb{Q}'$ and v, v is such that for almost all $x' \in \operatorname{supp}(\mathbb{Q}')$ there exists an $a \neq 0$ where for all $\varepsilon \in [0, |a|]$ it holds $x' + \varepsilon \operatorname{sign}(a) v \in \operatorname{supp}(\mathbb{Q}')$.

Since $f^* \in C^1(X)$ it holds

$$\frac{d}{d\varepsilon} f^*(x' + \varepsilon v)|_{\varepsilon = 0} = \langle v, \nabla_x f^*(x') \rangle.$$

Using Eq. 7 we see,

$$
\begin{aligned}
&\mathbb{E}_{x\sim\mathbb{P}}\left[\frac{f^*(x)-f^*(x'+\varepsilon v)}{\|x-(x'+\varepsilon v)\|}\right] \\
=&\varepsilon\left\langle v, \nabla_{\hat{x}}\mathbb{E}_{x\sim\mathbb{P}}\left[\frac{f^*(x)-f^*(\hat{x})}{\|x-\hat{x}\|}\right]\Big|_{x'}\right\rangle+\mathcal{O}(\varepsilon^2) \\
=&\frac{\mathbb{E}_{\tilde{x}\sim\mathbb{P}}\left[\left\langle \varepsilon(\tilde{x}-x'), \nabla_{\tilde{x}}\mathbb{E}_{x\sim\mathbb{P}}\left[\frac{f^*(x)-f^*(\hat{x})}{\|x-\hat{x}\|}\right]\Big|_{x'}\right\rangle\frac{f^*(\tilde{x})-f^*(x')}{\|x'-\tilde{x}\|^3}\right]}{\left\|\mathbb{E}_{\tilde{x}\sim\mathbb{P}}\left[(\tilde{x}-x')\frac{f^*(\tilde{x})-f^*(x')}{\|x'-\tilde{x}\|^3}\right]\right\|}+\mathcal{O}(\varepsilon^2) \\
=&\frac{\mathbb{E}_{\tilde{x}\sim\mathbb{P}}\left[\underbrace{\mathbb{E}_{x\sim\mathbb{P}}\left[\frac{f^*(x)-f^*(x'+\varepsilon(\tilde{x}-x'))}{\|x-x'+\varepsilon(\tilde{x}-x')\|}\right]}_{=\frac{1}{2\lambda}\text{, Eq. 7 and definition of }\mathbb{Q}'}\frac{f^*(\tilde{x})-f^*(x')}{\|x'-\tilde{x}\|^3}\right]}{\left\|\mathbb{E}_{\tilde{x}\sim\mathbb{P}}\left[(\tilde{x}-x')\frac{f^*(\tilde{x})-f^*(x')}{\|x'-\tilde{x}\|^3}\right]\right\|}+\mathcal{O}(\varepsilon^2) \\
=&\frac{\mathbb{E}_{\tilde{x}\sim\mathbb{P}}\left[\frac{1}{2\lambda}\frac{f^*(\tilde{x})-f^*(x')}{\|x'-\tilde{x}\|^3}\right]}{\left\|\mathbb{E}_{\tilde{x}\sim\mathbb{P}}\left[(\tilde{x}-x')\frac{f^*(\tilde{x})-f^*(x')}{\|x'-\tilde{x}\|^3}\right]\right\|}+\mathcal{O}(\varepsilon^2)
\end{aligned}
$$

which means

$$
\begin{aligned}
0=&\frac{d}{d\varepsilon}\mathbb{E}_{x\sim\mathbb{P}}\left[\frac{f^*(x)-f^*(x'+\varepsilon v)}{\|x-(x'+\varepsilon v)\|}\right]\Big|_{\varepsilon=0} \\
=&-\frac{d}{d\varepsilon}f^*(x'+\varepsilon v)|_{\varepsilon=0}\mathbb{E}_{\tilde{x}\sim\mathbb{P}}\left[\frac{1}{\|\tilde{x}-x'\|}\right]-\mathbb{E}_{x\sim\mathbb{P}}\left[\langle v, x-x'\rangle\frac{f^*(x)-f^*(x')}{\|x-x'\|^3}\right].
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\frac{d}{d\varepsilon}f^*(x'+\varepsilon v)|_{\varepsilon=0}=&\langle v, \nabla_x f^*(x)|_{x'}\rangle \\
=&\frac{\mathbb{E}_{\tilde{x}\sim\mathbb{P}}\left[\langle v, \tilde{x}-x'\rangle\frac{f^*(\tilde{x})-f^*(x')}{\|x-x'\|^3}\right]}{\mathbb{E}_{\tilde{x}\sim\mathbb{P}}\left[\frac{1}{\|\tilde{x}-x'\|}\right]} \\
=&\frac{\left\langle v, \mathbb{E}_{\tilde{x}\sim\mathbb{P}}[(\tilde{x}-x')\frac{f^*(\tilde{x})-f^*(x')}{\|x'-\tilde{x}\|^3}]\right\rangle}{\mathbb{E}_{\tilde{x}\sim\mathbb{P}}\left[\frac{1}{\|\tilde{x}-x'\|}\right]} \\
=&\frac{\left\|\mathbb{E}_{\tilde{x}\sim\mathbb{P}}[(\tilde{x}-x')\frac{f^*(\tilde{x})-f^*(x')}{\|x'-\tilde{x}\|^3}]\right\|}{\mathbb{E}_{\tilde{x}\sim\mathbb{P}}\left[\frac{1}{\|\tilde{x}-x'\|}\right]}
\end{aligned}
$$

Now from the proof of Lemma 5 claim (4), we know that since $G(\mathbb{P}, \mathbb{Q}; f^*)=0$ we get

$$
\|\nabla_x f^*(x)\big|_{x'}\|=\frac{\left\|\mathbb{E}_{\tilde{x}\sim\mathbb{P}}[(\tilde{x}-x')\frac{f^*(\tilde{x})-f^*(x')}{\|x'-\tilde{x}\|^3}]\right\|}{\mathbb{E}_{\tilde{x}\sim\mathbb{P}}[\frac{1}{\|x'-\tilde{x}\|}]}=\langle v, \nabla_x f^*(x)|_{x'}\rangle
$$

and since for $x\neq 0$ and $\|w\|=1$ it holds $\langle w, x\rangle=\|x\|\Leftrightarrow w\|x\|=x$ we discover $\nabla_x f^*(x)=v\|\nabla_x f^*(x)\big|_{x'}\|$ and thus

$$
\nabla_x f^*(x)\big|_{x'}=v\|\nabla_x f^*(x)\big|_{x'}\|=\frac{\mathbb{E}_{\tilde{x}\sim\mathbb{P}}\left[(\tilde{x}-x')\frac{f^*(\tilde{x})-f^*(x')}{\|x'-\tilde{x}\|^3}\right]}{\mathbb{E}_{\tilde{x}\sim\mathbb{P}}[\frac{1}{\|x'-\tilde{x}\|}]}
$$

and with

$$
\nabla_x f^*(x)\big|_{x'}\mathbb{E}_{\tilde{x}\sim\mathbb{P}}[\frac{1}{\|x'-\tilde{x}\|}]=\mathbb{E}_{\tilde{x}\sim\mathbb{P}}\left[(\tilde{x}-x')\frac{f^*(\tilde{x})-f^*(x')}{\|x'-\tilde{x}\|^3}\right].
$$

Plugging this into Eq. 14 gives us

$$\forall x' \in \mathrm{supp}(\mathbb{Q}') : \quad \nabla_x \mathbb{E}_{\tilde{x}\sim\mathbb{P}}\left[\frac{f^*(\tilde{x}) - f^*(x)}{\|\tilde{x} - x\|}\right]\Big|_{x'} = 0$$

□

Lemma 7. *Let $\mathbb{P}$ and $(\mathbb{Q}_\theta)_{\theta\in\Theta}$ in $\mathcal{P}(X)$ and fulfill Assumptions 1 and 2, further let $(\mathbb{Q}'_\theta)_{\theta\in\Theta}$ be as defined in introduction to Theorem 2, then for any $f^* \in \mathrm{OC}_{\tau_F}(\mathbb{P}, \mathbb{Q}'_\theta)$*

$$\nabla_\theta \tau_F(\mathbb{P}\|\mathbb{Q}'_\theta) \approx -\frac{1}{2}\nabla_\theta \mathbb{E}_{x'\sim\mathbb{Q}'_\theta}[f^*(x')]$$

thus f^* fulfills Eq. 5 and τ_F fulfills Requirement 4. Further, if $\mathbb{P}, \mathbb{Q}_\theta$ are such that there exits an f with $f(x) - f(x') = \|x - x'\|$ for all $x \in \mathrm{supp}(\mathbb{P})$ and $x' \in \mathrm{supp}(\mathbb{Q})$ then

$$\nabla_\theta \tau_F(\mathbb{P}\|\mathbb{Q}'_\theta) = -\frac{1}{2}\nabla_\theta \mathbb{E}_{x'\sim\mathbb{Q}'_\theta}[f^*(x')]$$

Proof. Start off by noting that for some $f^* \in \mathrm{OC}_{\tau_F}(\mathbb{P}, \mathbb{Q}_\theta)$, Theorem 1 from (Milgrom & Segal, 2002) gives us

$$\nabla_\theta \tau_F(\mathbb{P}\|\mathbb{Q}'_\theta)|_{\theta_0} = \nabla_\theta \tau_F(\mathbb{P}\|\mathbb{Q}'_\theta; f^*)|_{\theta_0}$$

Further, since for $f^* \in \mathrm{OC}_{\tau_F}(\mathbb{P}, \mathbb{Q}_\theta)$ it holds

$$\|\nabla_x f^*(x)\big|_{x'}\| = \frac{\left\|\mathbb{E}_{\tilde{x}\sim\mathbb{P}}[(\tilde{x} - x')\frac{f^*(\tilde{x})-f^*(x')}{\|x'-\tilde{x}\|^3}]\right\|}{\mathbb{E}_{\tilde{x}\sim\mathbb{P}}[\frac{1}{\|x'-\tilde{x}\|}]}$$

the gradient of the gradient penalty part is zero, i.e.

$$\nabla_\theta \mathbb{E}_{x\sim\mathbb{P}, x'\sim\mathbb{Q}_\theta}\left[\left(\|\nabla_x f^*(x)\big|_{x'}\| - \frac{\left\|\mathbb{E}_{\tilde{x}\sim\mathbb{P}}[(\tilde{x} - x')\frac{f^*(\tilde{x})-f^*(x')}{\|x'-\tilde{x}\|^3}]\right\|}{\mathbb{E}_{\tilde{x}\sim\mathbb{P}}[\frac{1}{\|x'-\tilde{x}\|}]}\right)^2\right] = 0.$$

One last point needs to be made before the main equation, which is for $x \in \mathrm{supp}(\mathbb{P})$

$$\nabla_\theta \mathbb{E}_{x'\sim\mathbb{Q}'_\theta}\left[\frac{f^*(x) - f^*(x')}{\|x - x'\|}\right] \approx 0.$$

This is from the motivation of the penalized Wasserstein GAN where for an optimal critic it should hold that $f^*(x) - f^*(x')$ is close to $c\|x - x'\|$ for some constant c. Note that if $\mathbb{P}$ and $\mathbb{Q}_\theta$ are such that $f^*(x) - f^*(x') = c\|x - x'\|$ is possible everywhere, then this term is equal to zero.

$$\nabla_\theta \tau_F(\mathbb{P}\|\mathbb{Q}'_\theta)|_{\theta_0} = \nabla_\theta \mathbb{E}_{\mathbb{P}\otimes\mathbb{Q}'_\theta}[(f^*(x) - f^*(x'))(1 - \lambda\frac{f^*(x) - f^*(x')}{\|x - x'\|})]|_{\theta_0}.$$

Since $\mathbb{Q}_\theta$ fulfills Assumption 1, $\mathbb{Q}_\theta \sim g(\theta, z)$ where g is differentiable in the first argument and $z \sim \mathbb{Z}$ ($\mathbb{Z}$ was defined in Assumption 1). Therefore if we set $g_\theta(\cdot) = g(\theta, \cdot)$ we get

$$\begin{aligned}
\nabla_\theta \tau_F(\mathbb{P}\|\mathbb{Q}')|_{\theta_0} &= \nabla_\theta \mathbb{E}_{x,\tilde{x}\sim\mathbb{P}, z\sim\mathbb{Z}, \alpha\sim\mathcal{U}([0,\varepsilon])}\left[(f^*(x) - f^*(\alpha\tilde{x} + (1-\alpha)g_\theta(z)))\left(1 - \lambda\frac{f^*(x) - f^*(\alpha\tilde{x} + (1-\alpha)g_\theta(z))}{\|x - \alpha\tilde{x} + (1-\alpha)g_\theta(z)\|}\right)\right]\Bigg|_{\theta_0} \\
&= -\mathbb{E}_{x,\tilde{x}\sim\mathbb{P}, z\sim\mathbb{Z}, \alpha\sim\mathcal{U}([0,\varepsilon])}\left[\nabla_\theta f^*(\alpha\tilde{x} + (1-\alpha)g_\theta(z))|_{\theta_0}\left(1 - \lambda\frac{f^*(x) - f^*(\alpha\tilde{x} + (1-\alpha)g_{\theta_0}(z))}{\|x - \alpha\tilde{x} + (1-\alpha)g_{\theta_0}(z)\|}\right)\right] \quad (17)\\
&\quad - \lambda\mathbb{E}_{x,\tilde{x}\sim\mathbb{P}, z\sim\mathbb{Z}, \alpha\sim\mathcal{U}([0,\varepsilon])}\left[(f^*(x) - f^*(\alpha\tilde{x} + (1-\alpha)g_{\theta_0}(z)))\nabla_\theta\left(\frac{f^*(x) - f^*(\alpha\tilde{x} + (1-\alpha)g_\theta(z))}{\|x - \alpha\tilde{x} + (1-\alpha)g_\theta(z)\|}\right)\Bigg|_{\theta_0}\right]. \quad (18)
\end{aligned}$$

Now if we look at the 17 term of the equation, we see that it's equal to:

$$
\begin{aligned}
&-\mathbb{E}_{\tilde{x}\sim\mathbb{P},z\sim\mathbb{Z},\alpha\sim\mathcal{U}([0,\varepsilon])}\left[\nabla_\theta f^*(\alpha\tilde{x}+(1-\alpha)g_\theta(z))|_{\theta_0}\underbrace{\mathbb{E}_{x\sim\mathbb{P}}\left[1-\lambda\frac{f^*(x)-f^*(\alpha\tilde{x}+(1-\alpha)g_{\theta_0}(z))}{\|x-\alpha\tilde{x}+(1-\alpha)g_{\theta_0}(z)\|}\right]}_{=\frac{1}{2}\text{, Eq. 7 from Lemma 2}}\right] \\
&= -\frac{1}{2}\nabla_\theta\mathbb{E}_{x'\sim\mathbb{Q}'_\theta}[f^*(x')]|_{\theta_0}
\end{aligned}
$$

and term 18 of the equation is equal to

$$
\begin{aligned}
&-\lambda\mathbb{E}_{x\sim\mathbb{P}}\left[f^*(x)\underbrace{\nabla_\theta\mathbb{E}_{x'\sim\mathbb{Q}'_\theta}\left[\frac{f^*(x)-f^*(x')}{\|x-x'\|}\right]}_{\approx 0\text{, See above}}\Bigg|_{\theta_0}\right] \\
&+\lambda\mathbb{E}_{\tilde{x}\sim\mathbb{P},z\sim\mathbb{Z},\alpha\sim\mathcal{U}([0,\varepsilon])}\left[f^*(\alpha\tilde{x}+(1-\alpha)g_{\theta_0}(z))\underbrace{\nabla_\theta\mathbb{E}_{x\sim\mathbb{P}}\left[1-\lambda\frac{f^*(x)-f^*(\alpha\tilde{x}+(1-\alpha)g_\theta(z))}{\|x-\alpha\tilde{x}+(1-\alpha)g_\theta(z)\|}\right]}_{=0\text{, Eq. 16}}\Bigg|_{\theta_0}\right]
\end{aligned}
$$

thus showing

$$
\nabla_\theta\tau_F(\mathbb{P}\|\mathbb{Q}'_\theta)|_{\theta_0}\approx-\frac{1}{2}\nabla_\theta\mathbb{E}_{x'\sim\mathbb{Q}'_\theta}[f^*(x')]|_{\theta_0}
$$

□

Lemma 8. *Let τ_I be the WGAN-GP divergence defined in Eq. 3, let the target distribution be the Dirac distribution δ_0 and the family of generated distributions be the uniform distributions $\mathcal{U}([0,\theta])$ with $\theta>0$. Then there is no $C\in\mathbb{R}$ that fulfills Eq. 5 for all $\theta>0$.*

Proof. For convenience, we'll restrict ourselves to the $\lambda=1$ case, for $\lambda\neq 1$ the proof is similar. Assume that $f\in \mathrm{OC}_{\tau_I}(\delta_0,\mathcal{U}([0,\theta])$ and $f(0)=0$. Since f is an optimal critic, for any function $u\in C^1(X)$ and any $\varepsilon\in\mathbb{R}$ it holds $\tau_I(\delta_0\|\mathcal{U}([0,\theta]);f)\geq\tau_I(\delta_0\|\mathcal{U}([0,\theta]);f+\varepsilon u)$. Therefore $\varepsilon=0$ is a maximum of the continuously differentiable function $\varepsilon\mapsto\tau_I(\delta_0\|\mathcal{U}([0,\theta]);f+\varepsilon u)$, and $\frac{d}{d\varepsilon}\tau_I(\delta_0\|\mathcal{U}([0,\theta]);f+\varepsilon u)|_{\varepsilon=0}=0$. Therefore

$$
\frac{d}{d\varepsilon}\tau_I(\delta_0\|\mathcal{U}([0,\theta]);f+\varepsilon u)|_{\varepsilon=0}=-\int_0^\theta u(t)dt-\int_0^\theta\frac{2}{t}\int_0^t u'(x)(f'(x)+1)dx\,dt=0
$$

multiplying by -1 and deriving with respect to θ gives us

$$
u(\theta)+\frac{2}{\theta}\int_0^\theta u'(x)(f'(x)+1)\,dx=0.
$$

Since we already made the assumption that $f(0)=0$ and since $\tau_I(\mathbb{P}\|\mathbb{Q};f)=\tau_I(\mathbb{P}\|\mathbb{Q};f+c)$ for any constant c, we can assume that $u(0)=0$. This gives us $u(\theta)=\int_0^\theta u'(x)\,dx$ and thus

$$
\int_0^\theta u'(x)\,dx+\frac{2}{\theta}\int_0^\theta u'(x)(f'(x)+1)\,dx=\frac{2}{\theta}\int_0^\theta u'(x)\left(\frac{\theta}{2}+f'(x)+1\right)\,dx.
$$

Therefore, for the optimal critic it holds $f'(x)=-(\frac{\theta}{2}+1)$, and since $f(0)=0$ the optimal critic is $f(x)=-(\frac{\theta}{2}+1)x$. Now

$$
\frac{d}{d\theta}\mathbb{E}_{\mathcal{U}([0,\theta])}[f]=-\frac{d}{d\theta}\int_0^\theta\left(\frac{\theta}{2}+1\right)x\,dx=-\left(\frac{\theta}{2}+1\right)\theta
$$

and

$$
\frac{d}{d\theta}\mathbb{E}_{\delta_0\otimes\mathcal{U}([0,\theta])}[r_f]=\frac{d}{d\theta}\frac{1}{\theta}\int_0^\theta\left(\frac{\theta}{2}\right)^2\,dx=\frac{d}{d\theta}\frac{\theta^2}{4}=\frac{\theta}{2}.
$$

Therefore there exists no $\gamma\in\mathbb{R}$ such that Eq. 5 holds for every distribution in the WGAN-GP context. □

B. Experiments

B.1. CelebA

The parameters used for CelebA training were:

```
'batch_size': 64,
'beta1': 0.5,
'c_dim': 3,
'calculate_slope': True,
'checkpoint_dir': 'logs/1127_220919_.0001_.0001/checkpoints',
'checkpoint_name': None,
'counter_start': 0,
'data_path': 'celebA_cropped/',
'dataset': 'celebA',
'discriminator_batch_norm': False,
'epoch': 81,
'fid_batch_size': 100,
'fid_eval_steps': 5000,
'fid_n_samples': 50000,
'fid_sample_batchsize': 1000,
'fid_verbose': True,
'gan_method': 'penalized_wgan',
'gradient_penalty': 1.0,
'incept_path': 'inception-2015-12-05/classify_image_graph_def.pb',
'input_fname_pattern': '*.jpg',
'input_height': 64,
'input_width': None,
'is_crop': False,
'is_train': True,
'learning_rate_d': 0.0001,
'learning_rate_g': 0.0005,
'lipschitz_penalty': 0.5,
'load_checkpoint': False,
'log_dir': 'logs/0208_191248_.0001_.0005/logs',
'lr_decay_rate_d': 1.0,
'lr_decay_rate_g': 1.0,
'num_discriminator_updates': 1,
'optimize_penalty': False,
'output_height': 64,
'output_width': None,
'sample_dir': 'logs/0208_191248_.0001_.0005/samples',
'stats_path': 'stats/fid_stats_celeba.npz',
'train_size': inf,
'visualize': False
```

The learned networks (both generator and critic) are then fine-tuned with learning rates divided by 10. Samples from the trained model can be viewed in figure 3.

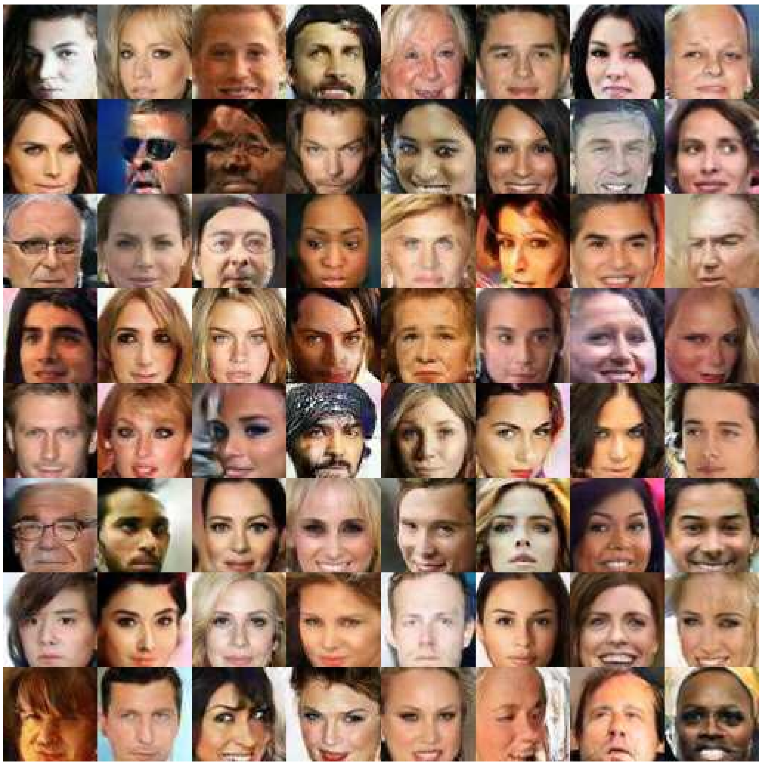

Figure 3. Images from a First Order GAN after training on CelebA data set.

B.2. CIFAR-10

The parameters used for CIFAR-10 training were:

```
BATCH_SIZE: 64
BETA1_D: 0.0
BETA1_G: 0.0
BETA2_D: 0.9
BETA2_G: 0.9
BN_D: True
BN_G: True
CHECKPOINT_STEP: 5000
CRITIC_ITERS: 1
DATASET: cifar10
DATA_DIR: /data/cifar10/
DIM: 32
D_LR: 0.0003
FID_BATCH_SIZE: 200
FID_EVAL_SIZE: 50000
FID_SAMPLE_BATCH_SIZE: 1000
FID_STEP: 5000
GRADIENT_PENALTY: 1.0
G_LR: 0.0001
INCEPTION_DIR: /data/inception-2015-12-05
ITERS: 500000
ITER_START: 0
LAMBDA: 10
LIPSCHITZ_PENALTY: 0.5
LOAD_CHECKPOINT: False
LOG_DIR: logs/
MODE: fogan
N_GPUS: 1
OUTPUT_DIM: 3072
OUTPUT_STEP: 200
SAMPLES_DIR: /samples
SAVE_SAMPLES_STEP: 200
STAT_FILE: /stats/fid_stats_cifar10_train.npz
TBOARD_DIR: /logs
TTUR: True
```

The learned networks (both generator and critic) are then fine-tuned with learning rates divided by 10. Samples from the trained model can be viewed in figure 4.

Figure 4. Images from a First Order GAN after training on CIFAR-10 data set.

B.3. LSUN

The parameters used for LSUN Bedrooms training were:

```
BATCH_SIZE: 64
 BETA1_D: 0.0
 BETA1_G: 0.0
 BETA2_D: 0.9
 BETA2_G: 0.9
 BN_D: True
 BN_G: True
 CHECKPOINT_STEP: 4000
 CRITIC_ITERS: 1
 DATASET: lsun
 DATA_DIR: /data/lsun
 DIM: 64
 D_LR: 0.0003
 FID_BATCH_SIZE: 200
 FID_EVAL_SIZE: 50000
 FID_SAMPLE_BATCH_SIZE: 1000
 FID_STEP: 4000
 GRADIENT_PENALTY: 1.0
 G_LR: 0.0001
 INCEPTION_DIR: /data/inception-2015-12-05
 ITERS: 500000
 ITER_START: 0
 LAMBDA: 10
 LIPSCHITZ_PENALTY: 0.5
 LOAD_CHECKPOINT: False
 LOG_DIR: /logs
 MODE: fogan
 N_GPUS: 1
 OUTPUT_DIM: 12288
 OUTPUT_STEP: 200
 SAMPLES_DIR: /samples
 SAVE_SAMPLES_STEP: 200
 STAT_FILE: /stats/fid_stats_lsun.npz
 TBOARD_DIR: /logs
 TTUR: True
```

The learned networks (both generator and critic) are then fine-tuned with learning rates divided by 10. Samples from the trained model can be viewed in figure 5.

Figure 5. Images from a First Order GAN after training on LSUN data set.

```
Change spent kands that the righ
Qust of orlists are mave hor int
Is that the spens has lought ant
If a took and their osiy south M
Willing contrased vackering in S
The Ireas's last to vising 5t ..
The FNF sicker , Nalnelber once
She 's wast to miblue as ganemat
threw pirnatures for hut only a
Umialasters are not oversup on t
Beacker it this that that that W
Though 's lunge plans wignsper c
He says : WalaMurka in the moroe
```

```
Dry Hall Sitning tven the concer
There are court phinchs hasffort
He scores a supponied foutver il
Bartfol reportings ane the depor
Seu hid , it 's watter 's remold
Later fasted the store the inste
Indiwezal deducated belenseous K
Starfers on Rbama 's all is lead
Inverdick oper , caldawho 's non
She said , five by theically rec
RichI , Learly said remain ."“
Reforded live for they were like
The plane was git finally fuels
```

Figure 6. Samples generated by First Order GAN trained on fhe One Billion Word benchmark with FOGAN (left) the original TTUR method (right).

B.4. Billion Word

The parameters used for the Billion Word training were one run with the following settings, followed by a second run using initialized with the best saved model from the first run and learning rates divided by 10. Samples from our method and the WGAN-GP baseline can be found in figure 6

```
 'activation_d': 'relu',
'batch_norm_d': False,
'batch_norm_g': True,
'batch_size': 64,
'checkpoint_dir': 'logs/checkpoints/0201_181559_0.000300_0.000100',
'critic_iters': 1,
'data_path': '1-billion-word-language-modeling-benchmark-r13output',
'dim': 512,
'gan_divergence': 'FOGAN',
'gradient_penalty': 1.0,
'is_train': True,
'iterations': 500000,
'jsd_test_interval': 2000,
'learning_rate_d': 0.0003,
'learning_rate_g': 0.0001,
'lipschitz_penalty': 0.1,
'load_checkpoint_dir': 'False',
'log_dir': 'logs/tboard/0201_181559_0.000300_0.000100',
'max_n_examples': 10000000,
'n_ngrams': 6,
'num_sample_batches': 100,
'print_interval': 100,
'sample_dir': 'logs/samples/0201_181559_0.000300_0.000100',
'seq_len': 32,
'squared_divergence': False,
'use_fast_lang_model': True
```