Appendix

A Proofs in Section 4

A.1 Proof of Lemma 1

By taking the second derivative of $f(\phi, \theta)$ in (3) w.r.t. ϕ, we have

$$\nabla^2_{\phi\phi} f(\phi, \theta) = \theta\theta^T \mathbb{E}_{z\sim\mathcal{N}(0,\sigma^2 I)}[g_2''(\theta^T(\phi + z))]$$

By the concavity of g_2, we know the scalar term $\mathbb{E}_{z\sim\mathcal{N}(0,\sigma^2 I)}[g_2''(\theta^T(\phi + z))] \leq 0$. Thus, we have

$$\nabla^2_{\phi\phi} f(\phi, \theta) \preceq 0$$

Similarly, by taking the second derivative of $f(\phi, \theta)$ in (3) w.r.t. θ, we have

$$\begin{aligned}\nabla^2_{\phi\phi} f(\phi, \theta) = {} & \mathbb{E}_{z\sim\mathcal{N}(v,\sigma^2 I)}[g_1''(\theta^T x) x x^T] \\ & + \mathbb{E}_{z\sim\mathcal{N}(0,\sigma^2 I)}[g_2''(\theta^T(\phi + z))(\phi + z)(\phi + z)^T]\end{aligned}$$

By the concavity of g_1 and g_2, we know the scalar terms $g_1''(\theta^T x) \leq 0$ and $g_2''(\theta^T(\phi + z)) \leq 0$. Since $xx^T \succeq 0$ and $(\phi + z)(\phi + z)^T \succeq 0$, we have

$$\nabla^2_{\theta\theta} f(\phi, \theta) \preceq 0$$

as required. □

A.2 Proof of Lemma 2

Proof. First, we have

$$\nabla_\phi f(\phi, \theta) = \mathbb{E}_{z\sim\mathcal{N}(0,\sigma^2 I)}[g_2'(\theta^T(\phi + z))\theta]$$

Since the equilibrium point (ϕ^*, θ^*) satisfies ${\theta^*}^T(\phi^* + z) = 0$, for points (ϕ, θ) near the equilibrium, we know $g_2'(\theta^T(\phi + z)) = g_2'(0) + g_2''(0)\theta^T(\phi + z) + o(\|\theta\|)$ by Taylor expansion. That is, by ignoring the small term with norm $o(\|\theta\|)$, we have

$$\begin{aligned}\nabla_\phi f(\phi, \theta) & \approx \mathbb{E}_{z\sim\mathcal{N}(0,\sigma^2 I)}[g_2'(0)\theta + g_2''(0)\theta\theta^T(\phi + z)] \\ & = g_2'(0)\theta + g_2''(0)\theta\theta^T\phi \\ & \overset{(a)}{\approx} g_2'(0)\theta\end{aligned}$$

where (a) is also from ignoring the small term with norm $o(\|\theta\|)$. Similarly,

$$\begin{aligned}\nabla_\theta f(\phi, \theta) = {} & \mathbb{E}_{x\sim\mathcal{N}(v,\sigma^2 I)}\left[g_1'(\theta^T x)x\right] \\ & + \mathbb{E}_{\tilde{x}\sim\mathcal{N}(\phi,\sigma^2 I)}\left[g_2'(\theta^T \tilde{x})\tilde{x}\right] \\ \overset{(a)}{\approx} & \mathbb{E}_{x\sim\mathcal{N}(v,\sigma^2 I)}\left[\left(g_1'(0) + g_1''(0)\theta^T x\right)x\right] \\ & + \mathbb{E}_{\tilde{x}\sim\mathcal{N}(\phi,\sigma^2 I)}\left[\left(g_2'(0) + g_2''(0)\theta^T \tilde{x}\right)\tilde{x}\right] \\ = & g_1'(0)v + g_1''(0)\left(\sigma^2 I + vv^T\right)\theta + g_2'(0)\phi \\ & + g_2''(0)\left(\sigma^2 I + \phi\phi^T\right)\theta \\ \overset{(b)}{\approx} & g_1'(0)v + g_2'(0)\phi + \left(g_1''(0) + g_2''(0)\right)\left(\sigma^2 I + vv^T\right)\theta\end{aligned}$$

where (a) is from $g_1'(\theta^T x) = g_1'(0) + g_1''(0)\theta^T x + o(\|\theta\|)$ and $g_2'(\theta^T \tilde{x}) = g_2'(0) + g_2''(0)\theta^T \tilde{x} + o(\|\theta\|)$ by Taylor expansion, and (b) is from $\|\phi - v\| = o(1)$.

For second-order derivatives, we have

$$\begin{aligned}\nabla^2_{\phi\phi} f(\phi, \theta) = & \mathbb{E}_{z\sim\mathcal{N}(0,\sigma^2 I)}\left[g_2''(\theta^T(\phi + z))\theta\theta^T\right] \\ \overset{(a)}{\approx} & g_2''(0)\theta\theta^T\end{aligned}$$

where (a) also follows from $g_2''(\theta^T(\phi + z)) = g_2''(0) + o(1)$ by Taylor expansion. Also,

$$\begin{aligned}\nabla^2_{\theta\phi} f(\phi, \theta) & = \mathbb{E}_{\tilde{x}\sim\mathcal{N}(\phi,\sigma^2 I)}[g_2'(\theta^T \tilde{x})I + g_2''(\theta^T \tilde{x})\tilde{x}\theta^T] \\ & \overset{(a)}{\approx} \mathbb{E}_{\tilde{x}\sim\mathcal{N}(\phi,\sigma^2 I)}[\left(g_2'(0) + g_2''(0)\theta^T \tilde{x}\right) I + g_2''(0)\tilde{x}\theta^T] \\ & = g_2'(0)I + g_2''(0)\theta^T\phi I + g_2''(0)\phi\theta^T \\ & \overset{(b)}{\approx} g_2'(0)I\end{aligned}$$

where (a) is from $g_2'(\theta^T \tilde{x}) = g_2'(0) + g_2''(0)\theta^T \tilde{x} + o(\|\theta\|)$ and $g_2''(\theta^T \tilde{x}) = g_2''(0) + o(1)$ by Taylor expansion, and (b) is from $\|\theta\| = o(1)$, and

$$\begin{aligned}\nabla^2_{\theta\theta} f(\phi, \theta) & = \mathbb{E}_{x\sim\mathcal{N}(v,\sigma^2 I)}\left[g_1''(\theta^T x)xx^T\right] \\ & \quad + \mathbb{E}_{\tilde{x}\sim\mathcal{N}(\phi,\sigma^2 I)}\left[g_2''(\theta^T \tilde{x})\tilde{x}\tilde{x}^T\right] \\ & \overset{(a)}{\approx} \mathbb{E}_{x\sim\mathcal{N}(v,\sigma^2 I)}\left[g_1''(0)xx^T\right] + \mathbb{E}_{\tilde{x}\sim\mathcal{N}(\phi,\sigma^2 I)}\left[g_2''(0)\tilde{x}\tilde{x}^T\right] \\ & \overset{(b)}{\approx} \left(g_1''(0) + g_2''(0)\right)\left(\sigma^2 I + vv^T\right)\end{aligned}$$

where (a) is from $g_1''(\theta^T x) = g_1''(0) + o(1)$ and $g_2''(\theta^T \tilde{x}) = g_2''(0) + o(1)$ by Taylor expansion, and (b) is from $\|\phi - v\| = o(1)$. □

A.3 Proof of Theorem 1

Proof. For the vanilla GAN, we know $g_1(t) = g_2(-t) = -\log(1 + e^{-t})$. Then we have $g_1'(0) = \frac{1}{2}$, $g_2'(0) = -\frac{1}{2}$ and $g_1''(0) = g_2''(0) = -\frac{1}{4}$. From the proof of Lemma 2, the updates (5) of SimGD for points near the equilibrium w^* become

$$\begin{aligned}w^{(k+1)} & = w^{(k)} + \eta \begin{bmatrix} \frac{1}{2}\theta^{(k)} \\ \frac{1}{2}(\phi^{(k)} - v) + \frac{1}{2}\left(\sigma^2 I + vv^T\right)\theta^{(k)} \end{bmatrix} \\ & = w^{(k)} + \eta \underbrace{\begin{bmatrix} 0 & \frac{1}{2}I \\ -\frac{1}{2}I & -\frac{1}{2}\left(\sigma^2 I + vv^T\right) \end{bmatrix}}_{\triangleq A} w^{(k)}\end{aligned} \tag{15}$$

where $w^{(k)} \triangleq \begin{bmatrix} \phi^{(k)} - v \\ \theta^{(k)} \end{bmatrix}$. Next, we need to compute the eigenvalues of the Jacobian A. By definition, let $Ay =$

λy where the eigenvector satisfies $y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \neq 0$, then we have

$$\frac{1}{2} y_2 = \lambda y_1 \tag{16}$$

$$-\frac{1}{2} y_1 - \frac{1}{2} \left(\sigma^2 I + v v^T \right) y_2 = \lambda y_2 \tag{17}$$

First, we know $\lambda \neq 0$, otherwise, we get $y = 0$ which violates the definition of eigenvectors. Thus from (16) we have $y_1 = \frac{1}{2\lambda} y_2$. Plugging it into (17) yields

$$-\lambda v v^T y_2 = (2\lambda^2 + \sigma^2 \lambda + \frac{1}{2}) y_2 \tag{18}$$

Then we can evaluate λ in two cases:

1) $v^T y_2 = 0$. From (18) we have $(4\lambda^2 + 2\sigma^2\lambda + 1) y_2 = 0$. Similarly we know $y_2 \neq 0$, otherwise, we get $y_1 = 0$ as well from (17) which again violates the definition of eigenvectors. Thus, the coefficient satisfies $4\lambda^2 + 2\sigma^2\lambda + 1 = 0$, and solving this equation yields $\lambda_{1,2}(A)$ in the theorem.

2) $v^T y_2 \neq 0$. By left multiplying v^T on both sides of Eq. (18) we get $-\lambda \|v\|^2 v^T y_2 = (2\lambda^2 + \sigma^2\lambda + \frac{1}{2}) v^T y_2$. Since $v^T y_2 \neq 0$, then $4\lambda^2 + 2(\sigma^2 + \|v\|^2)\lambda + 1 = 0$, and solving this equation yields $\lambda_{3,4}(A)$ in the theorem. □

A.4 Proof of Corollary 1

Proof. In the first part of the proof, we try to find the range of the step size η. Given $\sigma^2 < 2$, we know $\lambda_{1,2}(A)$ are complex eigenvalues and thus $|1 + \eta\lambda_{1,2}(A)| = \frac{1}{4}\eta^2 - \frac{\sigma^2}{2}\eta + 1$. Since it requires $|1 + \eta\lambda_{1,2}(A)| < 1$ to ensure the non-asymptotic convergence, by setting $\frac{1}{4}\eta^2 - \frac{\sigma^2}{2}\eta + 1 < 1$ we get $0 < \eta < 2\sigma^2$. As we know $\zeta = \sqrt{(\frac{2}{\sigma^2})^2 - 1}$ in the simple vanilla GAN example, then $\sigma^2 = \frac{2}{\sqrt{1+\zeta^2}}$, which means $0 < \eta < \frac{4}{\sqrt{1+\zeta^2}}$.

In the second part of the proof, we try to find the lower bound of the number of iterations N given the step size constraint. We know $\frac{1}{4}\eta^2 - \frac{\sigma^2}{2}\eta + 1 \geq \sqrt{1 - \left(\frac{\sigma^2}{2}\right)^2}$ with the equality holds at $\eta = \sigma^2$. Therefore, for the step size η satisfying $0 < \eta < \frac{4}{\sqrt{1+\zeta^2}}$, we have $\frac{1}{\sqrt{1+\frac{1}{\zeta^2}}} \leq |1 + \eta\lambda_{1,2}(A)| < 1$. Thus, for the updates $w^{(k)} = (I + \eta A) w^{(k-1)}$, it is easy to get $\tilde{w}^{(k)} = (I + \eta\Lambda)\tilde{w}^{(k-1)}$ where the eigen-matrix Λ satisfying $\Lambda = PAP^{-1}$ with P invertible and $\tilde{w}^{(k)} = Pw^{(k)}$. Apparently, $|\tilde{w}_j^{(k)}| = |I + \eta\lambda_{1,2}(A)|^k |\tilde{w}_j^{(0)}|$ where the index j refers to the entry in $\tilde{w}^{(k)}$ related to the eigenvalues $\lambda_{1,2}(A)$. Also, we know $\|\tilde{w}^{(k)}\| \geq |\tilde{w}_j^{(k)}|$ and $\|\tilde{w}^{(k)}\| = \|Pw^{(k)}\| \leq \|P\| \|w^{(k)}\|$, so we have $\|w^{(k)}\| \geq |I + \eta\lambda_{1,2}(A)|^k \|P\|^{-1} |\tilde{w}_j^{(0)}|$. Therefore, for the ϵ-error solution $\|w^{(N)}\| \leq \epsilon$ after N iterations, we have $(1 + \frac{1}{\zeta^2})^{-\frac{N}{2}} \|P\|^{-1} |\tilde{w}_j^{(0)}| \leq \epsilon$. By letting $C_0 = \|P\|^{-1} |\tilde{w}_j^{(0)}|$, we can easily get the lower bound of N. □

A.5 Proof of Corollary 2

Proof. In the first part of the proof, we try to find the range of the step size η. Given $\beta^2 > 2$, $\lambda_{3,4}(A)$ are both real eigenvalues. Similarly, to ensure the non-asymptotic convergence, the step size η also satisfies $|1 + \eta\lambda_{3,4}(A)| < 1$. From Theorem 1 we have $1 + \eta\lambda_3(A) = 1 - \frac{\beta^2 + \sqrt{(\beta^2)^2 - 4}}{4}\eta$ and $1 + \eta\lambda_4(A) = 1 - \frac{\beta^2 - \sqrt{(\beta^2)^2 - 4}}{4}\eta$. Next, we analyze $\lambda_3(A)$ and $\lambda_4(A)$ separately. To ensure $|1 + \eta\lambda_3(A)| < 1$, then $0 < \eta < \frac{8}{\beta^2 + \sqrt{(\beta^2)^2 - 4}}$. As we know $\tau = \frac{1}{4}(\beta^2 + \sqrt{(\beta^2)^2 - 4})^2$ in the simple vanilla GAN example, then $\frac{8}{\beta^2 + \sqrt{(\beta^2)^2 - 4}} = \frac{4}{\sqrt{\tau}}$, which means $0 < \eta < \frac{4}{\sqrt{\tau}}$. Also, to satisfy $|1 + \eta\lambda_4(A)| < 1$, then $0 < \eta < 2(\beta^2 + \sqrt{(\beta^2)^2 - 4}) = 4\sqrt{\tau}$. As we know $\tau > 1$ by definition, the step size η satisfies $0 < \eta < \min\{\frac{4}{\sqrt{\tau}}, 4\sqrt{\tau}\} = \frac{4}{\sqrt{\tau}}$.

In the second part of the proof, we try to find the lower bound of the number of iterations N given the step size constraint. We know $|1 + \eta\lambda_4(A)| = |1 - \frac{1}{2\sqrt{\tau}}\eta|$ and for $0 < \eta < \frac{4}{\sqrt{\tau}}$ we get $1 - \frac{2}{\tau} < 1 - \frac{1}{2\sqrt{\tau}}\eta < 1$, Therefore, if $1 < \tau < 2$, then $-1 < 1 - \frac{2}{\tau} < 0$, and thus $0 < |1 + \eta\lambda_4(A)| < 1$. If $\tau \geq 2$, then $1 - \frac{2}{\tau} > 0$, and thus $1 - \frac{2}{\tau} < |1 + \eta\lambda_4(A)| < 1$. Putting them together, we get $\max\{1 - \frac{2}{\tau}, 0\} < |1 + \eta\lambda_4(A)| < 1$. Similar to the proof of Corollary 1, we rewrite the updates as $\tilde{w}^{(k)} = (I + \eta\Lambda)\tilde{w}^{(k-1)}$ where the eigen-matrix Λ satisfying $\Lambda = PAP^{-1}$ with P invertible and $\tilde{w}^{(k)} = Pw^{(k)}$. Here we focus on $|\tilde{w}_{j'}^{(k)}| = |I + \eta\lambda_{1,2}(A)|^k |\tilde{w}_{j'}^{(0)}|$ where the index j' refers to the entry in $\tilde{w}^{(k)}$ related to the eigenvalues $\lambda_4(A)$. Also, we know $\|w^{(k)}\| \geq |I + \eta\lambda_{1,2}(A)|^k C_1$ where $C_1 = \|P\|^{-1} |\tilde{w}_{j'}^{(0)}|$. Therefore, for $\tau > 2$, we get $\|w^{(k)}\| \geq (1 - \frac{2}{\tau})^{\frac{k}{2}} C_1$. For the ϵ-error solution $\|w^{(N)}\| \leq \epsilon$, we have $(1 - \frac{2}{\tau})^{\frac{N}{2}} C_1 \leq \epsilon$ which yields the lower bound of N. □

B An Example of Full Rank Representations

In the simple vanilla GAN example, if we consider the zero noise-limit case, i.e. $\sigma^2 = 0$, and assume $n = 1$, from Theorem 1 we know the eigenvalues of the Jacobian

A are

$$\lambda_{1,2}(A) = \frac{-v^2 \pm \sqrt{\left(v^2\right)^2 - 4}}{4} \tag{19}$$

When $v \to 0$, $\lambda_{1,2}(A) \to \pm\frac{1}{2}i$ with an infinitely large imaginary-to-real ratio ζ, which obviously suffers from the impact of the *Phase Factor*.

To alleviate this issue, one solution could be to increase the expressive power of discriminator. For instance, it is suggested by Mescheder et al. (2018) that we can replace the linear discriminator $D_\theta(x) = \theta x$ by the discriminator with the so-called full-rank representations $D_\theta(x) = \theta e^x$. Similarly, in the zero noise-limit case with $n = 1$, we first rewrite the objective (3) as $f(\theta, \phi) = g_1(\theta e^x) + g_2(-\theta e^x)$. For the vanilla GAN, we have $g_1(t) = g_2(-t) = -\log(1 + e^{-t})$. Then the Jacobian A of all points within $B_\delta(w^*)$ is evaluated as $A = \begin{bmatrix} 0 & \frac{1}{2}e^v \\ -\frac{1}{2}e^v & -\frac{1}{2}e^{2v} \end{bmatrix}$ and its eigenvalues are

$$\lambda_{1,2}(A) = \frac{-e^{2v} \pm \sqrt{e^{4v} - 4e^{2v}}}{4} \tag{20}$$

Now when $v \to 0$, $\lambda_{1,2}(A) \to \frac{-1 \pm \sqrt{3}i}{4}$ with the imaginary-to-real ratio $\zeta = \sqrt{3}$. By Corollary 1, the impact of the *Phase Factor* has been effectively alleviated when v is very small.

However, the impact of the *Conditioning Factor*, if it exists, becomes much more severe. Asymptotically when v is sufficiently large, from (19) we know that τ increases in the order of v^4, but (20) shows that τ increases in the order of e^{2v}. For example, if we assume $v = 5$, the eigenvalues of the original Jacobian (19) is evaluated as $\lambda_{1,2}(A) = \frac{-25 \pm \sqrt{621}}{2}$ with $\tau = \Omega(10^2)$. However, after using the discriminator with full-rank representations, the eigenvalues of the new Jacobian (20) is evaluated as $\lambda_{1,2}(A) = \frac{-e^{10} \pm \sqrt{e^{20} - 4e^{10}}}{4}$ with $\tau = \Omega(10^5)$.

C A Condition of Choosing the Regularization Matrix

First, we note that the regularization matrix Γ introduced by a good Jacobian regularization method cannot be arbitrary and a particular condition is given as follows.

Condition 1 (Non-Reversing-Flow Condition). *By applying the regularization matrix Γ, it should not reverse the overall gradient flow for the original minimax problem (1).*

A counterexample of the Non-Reversing-Flow Condition is to choose $\Gamma = -M^T$ where $M \triangleq \frac{\partial \tilde{\nabla} f(w^{(k)})}{\partial w^{(k)}}^T$ such that the new Jacobian becomes $A = -M^T M$. Now the Jacobian A is a Hessian which has no complex eigenvalues and thus it could avoid the *Phase Factor*. From (6), the updates become

$$\begin{aligned} w^{(k+1)} &= w^{(k)} - \eta M^T \tilde{\nabla} f(w^{(k)}) \\ &= w^{(k)} - \eta \nabla^2 f(w^{(k)}) \nabla f(w^{(k)}) \end{aligned}$$

As we know, in general, the objective $f(\phi, \theta)$ is not convex-concave in ϕ and θ. For example, $f(\phi, \theta)$ becomes concave-concave in ϕ and θ near the equilibrium in the simple vanilla GAN example (3). Therefore, for any $w^{(k)}$ satisfying $\nabla^2_{\phi\phi} f(w^{(k)}) \prec 0$, particularly if assuming $\nabla^2_{\phi\phi} f(w^{(k)}) = -t^2 I$ where t is a non-zero scalar, the update for ϕ becomes

$$\phi^{(k+1)} = \phi^{(k)} + \eta t^2 \nabla_\phi f(w^{(k)}) - \eta \nabla^2_{\theta\phi} f(w^{(k)}) \nabla_\theta f(w^{(k)})$$

According to the first two terms on the right-hand side of the above equation, it is actually a gradient flow of the generator G_ϕ maximizing the objective $f(\phi, \theta)$ instead. This partly explains why directly minimizing a surrogate loss $l(w) = \frac{1}{2}\|\nabla f(w^{(k)})\|^2$ does not work well in practice as has been observed by Mescheder et al. (2017).

Next, we point out that ConOpt may also violate the Non-Reversing-Flow Condition in some cases. Similarly, for any point $w^{(k)}$ satisfying $\nabla^2_{\phi\phi} f(w^{(k)}) \prec 0$, particularly if we assume $\nabla^2_{\phi\phi} f(w^{(k)}) = -t^2 I$, the update for ϕ in (28) for ConOpt becomes

$$\begin{aligned} \phi^{(k+1)} =& \phi^{(k)} + \eta(\gamma t^2 - 1) \nabla_\phi f(w^{(k)}) \\ &- \eta\gamma \nabla^2_{\theta\phi} f(w^{(k)}) \nabla_\theta f(w^{(k)}) \end{aligned}$$

If $\gamma t^2 > 1$, it is also a gradient flow of the generator G_ϕ maximizing the objective $f(\phi, \theta)$ instead. Note that the Hessian $\nabla^2_{\phi\phi} f(w^{(k)})$, introduced by ConOpt to the parameter updates, serves as the root cause of violating Condition 1. This might also partly explains why ConOpt is less robust than our proposed method in some experiments. Even worse, as γ increases, it is more likely for ConOpt to reverse the gradient flow. It intuitively explains why γ should be kept relatively small for ConOpt.

D Proofs in Section 5

D.1 Proof of Theorem 2

Proof. we revisit each of these three regularization methods by evaluating and analyzing the eigenvalues of their Jacobians in the simple vanilla GAN example separately.

Only regularizing generator. The regularized updates for generator become

$$\phi^{(k+1)} = \phi^{(k)} - \eta \nabla_\phi f(w^{(k)}) - \frac{1}{2}\eta\gamma \nabla_\phi \left\| \nabla_\theta f(w^{(k)}) \right\|^2 \tag{21}$$

In the simple vanilla GAN example, from (4) in Lemma 2, $\frac{\partial \tilde{\nabla} f(w^{(k)})}{\partial w^{(k)}} = \begin{bmatrix} 0 & -\frac{1}{2}I \\ \frac{1}{2}I & -\frac{1}{2}\left(\sigma^2 I + vv^T\right) \end{bmatrix}$. Also the regularization matrix becomes $\Gamma = \begin{bmatrix} I & \frac{\gamma}{2}I \\ 0 & I \end{bmatrix}$. Thus, for all points in $B_\delta(w^*)$, the Jacobian is

$$\begin{aligned} A &= \Gamma \frac{\partial \tilde{\nabla} f(w^{(k)})}{\partial w^{(k)}}^T \\ &= \begin{bmatrix} I & \frac{\gamma}{2}I \\ 0 & I \end{bmatrix} \begin{bmatrix} 0 & \frac{1}{2}I \\ -\frac{1}{2}I & -\frac{1}{2}\left(\sigma^2 I + vv^T\right) \end{bmatrix} \\ &= \begin{bmatrix} -\frac{\gamma}{4}I & \frac{1}{2}I - \frac{\gamma}{4}\left(\sigma^2 I + vv^T\right) \\ -\frac{1}{2}I & -\frac{1}{2}\left(\sigma^2 I + vv^T\right) \end{bmatrix} \end{aligned}$$

By definition of eigenvalues, let $Ay = \lambda y$ where $y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \neq 0$, then

$$-\frac{\gamma}{4}y_1 + \left(\frac{1}{2}I - \frac{\gamma}{4}(\sigma^2 I + vv^T)\right) y_2 = \lambda y_1 \quad (22)$$

$$-\frac{1}{2}y_1 - \frac{1}{2}\left(\sigma^2 I + vv^T\right) y_2 = \lambda y_2 \quad (23)$$

From (22) we have $y_1 = \frac{1}{\lambda + \frac{\gamma}{4}}\left(\frac{1}{2}I - \frac{\gamma}{4}(\sigma^2 I + vv^T)\right) y_2$ (note that $\lambda \neq -\frac{\gamma}{4}$; otherwise, we get $y = 0$). Plugging it into (23) yields

$$-\lambda vv^T y_2 = \left(2\lambda^2 + (\frac{1}{2}\gamma + \sigma^2)\lambda + \frac{1}{2}\right) y_2 \quad (24)$$

Similarly, we can also solve (24) in two cases yielding the eigenvalues of the Jacobian as follows,

$$\begin{aligned} \lambda_{1,2}(A) &= \frac{-\left(\sigma^2 + \frac{\gamma}{2}\right) \pm \sqrt{\left(\sigma^2 + \frac{\gamma}{2}\right)^2 - 4}}{4}, \\ \lambda_{3,4}(A) &= \frac{-\left(\beta^2 + \frac{\gamma}{2}\right) \pm \sqrt{\left(\beta^2 + \frac{\gamma}{2}\right)^2 - 4}}{4} \end{aligned} \quad (25)$$

As we can see, the resulting $\zeta = \sqrt{(\frac{2}{\sigma^2 + \frac{\gamma}{2}})^2 - 1}$ for $\sigma^2 + \frac{\gamma}{2} < 2$, which means increasing γ will decrease ζ and thus could alleviate the impact of the *Phase Factor* by Corollary 1. However, the resulting $\tau = \frac{\left((\beta^2 + \frac{\gamma}{2}) + \sqrt{(\beta^2 + \frac{\gamma}{2})^2 - 4}\right)^2}{4}$ for $\beta^2 + \frac{\gamma}{2} > 2$, which means increasing γ will also increase τ and thus the impact of *Conditioning Factor* will not be alleviated but become much severer by Corollary 2. Therefore, if the *Conditioning Factor* is the main obstacle for the GAN convergence (for example, $\|v\|$ is sufficiently large in the simple vanilla GAN example), only regularizing generator as in (21) will make the convergence performance of the GAN training worse.

Only regularizing discriminator. The regularized updates for the discriminator become

$$\theta^{(k+1)} = \theta^{(k)} + \eta \nabla_\theta f(w^{(k)}) - \frac{1}{2}\eta\gamma\nabla_\theta \left\|\nabla_\phi f(w^{(k)})\right\|^2 \quad (26)$$

Similarly in the simple vanilla GAN example, the regularziation matrix becomes $\Gamma = \begin{bmatrix} I & 0 \\ -\frac{\gamma}{2}I & I \end{bmatrix}$. For any point in $B_\delta(w^*)$, the Jacobian is

$$\begin{aligned} A &= \Gamma \frac{\partial \tilde{\nabla} f(w^{(k)})}{\partial w^{(k)}}^T \\ &= \begin{bmatrix} I & 0 \\ -\frac{\gamma}{2}I & I \end{bmatrix} \begin{bmatrix} 0 & \frac{1}{2}I \\ -\frac{1}{2}I & -\frac{1}{2}\left(\sigma^2 I + vv^T\right) \end{bmatrix} \\ &= \begin{bmatrix} 0 & \frac{1}{2}I \\ -\frac{1}{2}I & -\frac{1}{2}\left(\left(\sigma^2 + \frac{\gamma}{2}\right) I + vv^T\right) \end{bmatrix} \end{aligned}$$

Then by following from the exact proof of Theorem 1 after replacing σ^2 in the Jacobian of (15) by $\sigma^2 + \frac{\gamma}{2}$, we can get the eigenvalues of the Jacobian as follows,

$$\begin{aligned} \lambda_{1,2}(A) &= \frac{-\left(\sigma^2 + \frac{\gamma}{2}\right) \pm \sqrt{\left(\sigma^2 + \frac{\gamma}{2}\right)^2 - 4}}{4}, \\ \lambda_{3,4}(A) &= \frac{-\left(\beta^2 + \frac{\gamma}{2}\right) \pm \sqrt{\left(\beta^2 + \frac{\gamma}{2}\right)^2 - 4}}{4} \end{aligned} \quad (27)$$

As the eigenvalues here are exactly the same with (25), the local convergence properties of only regularizing the discriminator are identical to those of only regularizing the generator. Similarly, if *Conditioning Factor* becomes the main obstacle for GAN convergence, only regularizing discriminator as in (26) will make the convergence performance of the GAN training worse.

Consensus optimization (ConOpt). The regularized updates for the generator and discriminator are

$$w^{(k+1)} = w^{(k)} + \eta \tilde{\nabla} f(w^{(k)}) - \frac{1}{2}\eta\gamma\nabla \left\|\nabla f(w^{(k)})\right\|^2 \quad (28)$$

Since for ConOpt, it is a little bit tricky to obtain the eigenvalues of its Jacobian directly, we turn to comparing the eigenvalues of it Jacobian with those of the Jacobian for SimGD.

First, we define $M \triangleq \frac{\partial \tilde{\nabla} f(w^{(k)})}{\partial w^{(k)}}^T$. For SimGD, we know its Jacobian is M. For ConOpt, since the regularization matrix $\Gamma = I - \gamma M^T$, its Jacobian is

$$A = \Gamma M = M - \gamma M^T M \quad (29)$$

Then, we define $\overline{\lambda}(M)$ and $\underline{\lambda}(M)$ as the two eigenvalues of M with the largest and smallest absolute values, respectively, and the similar definitions of $\overline{\lambda}(A)$ and $\underline{\lambda}(A)$

apply to A. Thus, the condition numbers of A and M are $\tau(A) \triangleq \frac{|\overline{\lambda}(A)|}{|\underline{\lambda}(A)|}$ and $\tau(M) \triangleq \frac{|\overline{\lambda}(M)|}{|\underline{\lambda}(M)|}$, respectively.

If $\sigma^2 < 2$ and $\beta^2 > 2$, from Theorem 1 we know for any point in $B_\delta(w^*)$, the Jacobian for SimGD satisfies $|\lambda_{1,2}(M)| = \frac{1}{2}$, $|\lambda_3(M)| = \frac{\beta^2+\sqrt{(\beta^2)^2-4}}{4} > \frac{1}{2}$ and $|\lambda_4(M)| = \frac{\beta^2-\sqrt{(\beta^2)^2-4}}{4} < \frac{1}{2}$. Thus, $\overline{\lambda}(M) = \lambda_3(M)$ and $\underline{\lambda}(M) = \lambda_4(M)$, which are both negative values.

By definition of eigenvalues, we have $My_1 = \overline{\lambda}(M)y_1$ and $My_2 = \underline{\lambda}(M)y_2$ where y_1 and y_2 are two normalized eigenvectors of M with unit length. Thus, $y_1^T My_1 = \overline{\lambda}(M)$ and $y_2^T My_2 = \underline{\lambda}(M)$. From (29), we have $y_1^T My_1 = \overline{\lambda}(M) - \gamma\overline{\lambda}(M)^2$ and $y_2^T My_2 = \underline{\lambda}(M) - \gamma\underline{\lambda}(M)^2$. From the definition of $\overline{\lambda}(A)$ and $\underline{\lambda}(A)$, we know $|y_1^T Ay_1| \leq |\overline{\lambda}(A)|$ and $|y_2^T Ay_2| \geq |\underline{\lambda}(A)|$, then $|\overline{\lambda}(M) - \gamma\overline{\lambda}(M)^2| \leq |\overline{\lambda}(A)|$ and $|\underline{\lambda}(M) - \gamma\underline{\lambda}(M)^2| \geq |\underline{\lambda}(A)|$. Combining the two inequalities yields

$$\tau(A) \geq \tau(M) \cdot \frac{1+\gamma|\overline{\lambda}(M)|}{1+\gamma|\underline{\lambda}(M)|} \tag{30}$$

Define by $\Delta(\gamma) \triangleq \frac{1+\gamma|\overline{\lambda}(M)|}{1+\gamma|\underline{\lambda}(M)|}$. As $|\overline{\lambda}(M)| > |\underline{\lambda}(M)| > 0$ and $\gamma > 0$, we have $\Delta(\gamma) > 1$, which means $\tau(A) > \tau(M)$ for any $\gamma > 0$. Even worse, since the derivative $\Delta'(\gamma) = \frac{\overline{\lambda}(M)-\underline{\lambda}(M)}{(1-\gamma\underline{\lambda}(M))^2} > 0$, when γ increases, $\Delta(\gamma)$ also increases. Thus, by using ConOpt, the impact of *Conditioning Factor* is not alleviated but becomes more severe by Corollary 2. Furthermore, the Jacobian will be worse-conditioned as γ increases. Therefore, although ConOpt could alleviate the impact of the *Phase Factor* as shown in Mescheder et al. (2017), it will make the GAN convergence performance worse if the *Conditioning Factor* becomes the main obstacle for the GAN convergence.

From the above analysis, all these three gradient-based regularization methods cannot alleviate the *Phase Factor* and *Conditioning Factor* simultaneously. □

D.2 Proof of Theorem 3

Proof. When applying the proposed Jacobian regularization in the simple vanilla GAN example (3), the regularization matrix becomes $\Gamma = \begin{bmatrix} I & \frac{\gamma}{2}I \\ -\frac{\gamma}{2}I & I \end{bmatrix}$. Therefore, for any point in $B_\delta(w^*)$,

$$\begin{aligned} A &= \Gamma \frac{\partial \tilde{\nabla} f(w^{(k)})}{\partial w^{(k)}}^T \\ &= \begin{bmatrix} I & \frac{\gamma}{2}I \\ -\frac{\gamma}{2}I & I \end{bmatrix} \begin{bmatrix} 0 & \frac{1}{2}I \\ -\frac{1}{2}I & -\frac{1}{2}\left(\sigma^2 I + vv^T\right) \end{bmatrix} \\ &= \begin{bmatrix} -\frac{\gamma}{4}I & \frac{1}{2}I - \frac{\gamma}{4}\left(\sigma^2 I + vv^T\right) \\ -\frac{1}{2}I & -\frac{\gamma}{4}I - \frac{1}{2}\left(\sigma^2 I + vv^T\right) \end{bmatrix} \end{aligned}$$

By definition of eigenvalues, let $Ay = \lambda y$ where $y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \neq 0$, then

$$-\frac{\gamma}{4}y_1 + \left(\frac{1}{2}I - \frac{\gamma}{4}(\sigma^2 I + vv^T)\right)y_2 = \lambda y_1 \tag{31}$$

$$-\frac{1}{2}y_1 - \frac{\gamma}{4}y_2 - \frac{1}{2}\left(\sigma^2 I + vv^T\right)y_2 = \lambda y_2 \tag{32}$$

Similarly, $\lambda \neq 0$, otherwise, we get $y = 0$ which violates the definition of eigenvectors. By applying $(31) - (32) * \frac{\gamma}{2}$, we have $y_1 = \frac{1}{\lambda}\left(\frac{\gamma}{2}\lambda + \frac{\gamma^2}{8} + \frac{1}{2}\right)y_2$. Plugging it into (32) yields

$$-\lambda vv^T y_2 = \left(2\lambda^2 + (\gamma + \sigma^2)\lambda + \frac{\gamma^2}{8} + \frac{1}{2}\right)y_2 \tag{33}$$

Similarly, we can solve (33) in two cases yielding the desired results by following the same process in the proof of Theorem 1. □

D.3 Proof of Corollary 3

Proof. From Theorem 3 we know for $\sigma^2 < 2$, $\lambda_{1,2}(A)$ are complex eigenvalues only if $\gamma < \frac{2}{\sigma^2} - \frac{\sigma^2}{2}$. According to the above definition of ζ, we get

$$\zeta = \begin{cases} \sqrt{h_1(\gamma) - 1}, & \gamma < \frac{2}{\sigma^2} - \frac{\sigma^2}{2} \\ 0, & \gamma \geq \frac{2}{\sigma^2} - \frac{\sigma^2}{2} \end{cases} \tag{34}$$

where $h_1(\gamma) = \frac{\gamma^2+4}{(\sigma^2+\gamma)^2} > 1$. Since the derivative of $h_1(\gamma)$ satisfies $h_1'(\gamma) = \frac{2(\gamma+\sigma^2)(\sigma^2\gamma-4)}{(\sigma^2+\gamma)^4} < 0$ and ζ is a monotonically increasing function of $h_1(\gamma)$ for $\gamma < \frac{2}{\sigma^2} - \frac{\sigma^2}{2}$, ζ is a monotonically decreasing function of γ for $\gamma < \frac{2}{\sigma^2} - \frac{\sigma^2}{2}$. As $\zeta = 0$ if $\gamma \geq \frac{2}{\sigma^2} - \frac{\sigma^2}{2}$, by the continuity of the function in (34), we have ζ is a monotonically decreasing function of γ where $\zeta \to 0$ as $\gamma \to \infty$. It means that we can increase γ to alleviate the impact of the *Phase Factor*.

Furthermore, from Theorem 3 we know for $\beta^2 > 2$,

$$\tau = \left(\sqrt{h_2(\gamma)} + \sqrt{h_2(\gamma)^2 - 1}\right)^2 \tag{35}$$

where $h_2(\gamma) = \frac{(\beta^2+\gamma)^2}{\gamma^2+4} > 1$. Since the derivative of $h_2(\gamma)$ satisfies $h_2'(\gamma) = \frac{2(\gamma+\beta^2)(4-\beta^2\gamma)}{(\gamma+4)^4} < 0$ for $\gamma > \frac{4}{\beta^2}$ and τ is a monotonically increasing function of $h_2(\gamma)$, τ is a monotonically decreasing function of γ for $\gamma > \frac{4}{\beta^2}$. As $\beta^2 > 2$, then $\frac{4}{\beta^2} < 2$ and we thus can safely replace the above condition $\gamma > \frac{4}{\beta^2}$ by $\gamma \geq 2$. In the limit of $\gamma \to \infty$, we have $h_2(\gamma) \to 1$ and thus from (35) $\tau \to 1$. It means that we can increase γ to alleviate the impact of the *Conditioning Factor* for all $\gamma > \frac{4}{\beta^2}$.

Therefore, it is reasonable to keep increasing the tunable parameter γ so as to alleviate or even eliminate both the *Phase Factor* and *Conditioning Factor* simultaneously, which demonstrates the advantages of JARE. □

E Proof in Section 6

E.1 Proof of Lemma 3

Although the proof is very similar to Mescheder et al. (2018), we provide the proof details for completeness.

Since we know the objective is

$$f(\phi, \theta) \triangleq \mathbb{E}_{x\sim P_r}[g_1(D_\theta(x))] + \mathbb{E}_{z\sim P_0}[g_2(D_\theta(G_\phi(z)))]$$

By taking its derivative w.r.t. ϕ and θ at the equilibrium (ϕ^*, θ^*), respectively, we have

$$\begin{aligned}\nabla_\phi f(\phi^*, \theta^*) = \mathbb{E}_{z\sim P_0}[&g_2'(D_{\theta^*}(x))\nabla_\phi G_{\phi^*}(z) \\ &\cdot \nabla_x D_{\theta^*}(x)]|_{x=G_{\phi^*}(z)}\end{aligned} \tag{36}$$

$$\begin{aligned}\nabla_\theta f(\phi^*, \theta^*) =&\mathbb{E}_{x\sim P_r}[g_1'(D_{\theta^*}(x))\nabla_\theta D_{\theta^*}(x)] \\ &+ \mathbb{E}_{x\sim P_{\phi^*}}[g_2'(D_{\theta^*}(x))\nabla_\theta D_{\theta^*}(x)]\end{aligned} \tag{37}$$

Since the Jacobian A at (ϕ^*, θ^*) in general GANs trained via SimGD are given by

$$A = \begin{bmatrix} -\nabla^2_{\phi\phi} f(\phi^*, \theta^*) & -\nabla^2_{\phi\theta} f(\phi^*, \theta^*) \\ \nabla^2_{\theta\phi} f(\phi^*, \theta^*) & \nabla^2_{\theta\theta} f(\phi^*, \theta^*) \end{bmatrix}$$

First, from Assumption 1 we know that $D_{\theta^*}(x) = 0$ for some local neighborhood of any $x \in \mathcal{X}$, which means we also have $\nabla_x D_{\theta^*}(x) = 0$ and $\nabla^2_{xx} D_{\theta^*}(x) = 0$ for any $x \in \mathcal{X}$. By taking the derivative of (36) w.r.t. ϕ at the equilibrium (ϕ^*, θ^*) and using $\nabla_x D_{\theta^*}(x) = 0$ and $\nabla^2_{xx} D_{\theta^*}(x) = 0$ for any $x \in \mathcal{X}$, we have

$$\nabla^2_{\phi\phi} f(\phi^*, \theta^*) = 0$$

By taking the derivative of (37) w.r.t. ϕ at the equilibrium (ϕ^*, θ^*), we have

$$\begin{aligned}\nabla^2_{\phi\theta} f(\phi^*, \theta^*) &= \mathbb{E}_{z\sim P_0}[g_2''(D_{\theta^*}(x))\nabla_\phi G_{\phi^*}(z) \\ &\qquad \cdot \nabla^2_{x\theta} D_{\theta^*}(x)]|_{x=G_{\phi^*}(z)} \\ &\overset{(a)}{=} g_2''(0)\mathbb{E}_{z\sim P_0}[\nabla_\phi G_{\phi^*}(z)\nabla^2_{x\theta} D_{\theta^*}(x)]|_{x=G_{\phi^*}(z)}\end{aligned}$$

where (a) is from the assumption that $D_{\theta^*} = 0$.

By taking the derivative of (37) w.r.t. θ, respectively, at the equilibrium (ϕ^*, θ^*), we have

$$\begin{aligned}\nabla^2_{\theta\theta} f(\phi^*, \theta^*) \overset{(a)}{=}&\mathbb{E}_{x\sim P_r}[(g_1'(0) + g_2'(0))\nabla^2_{\theta\theta} D_{\theta^*}(x) \\ &+ (g_1''(0) + g_2''(0))\nabla_\theta D_{\theta^*}(x) D_{\theta^*}(x)^T] \\ \overset{(b)}{=}&(g_1''(0) + g_2''(0))\mathbb{E}_{x\sim P_r}[\nabla_\theta D_{\theta^*}(x) D_{\theta^*}(x)^T]\end{aligned}$$

where (a) is from Assumption 1 that $P_r = P_{\phi^*}$ and $D_{\theta^*} = 0$, (b) is from Assumption 2 that $g_1'(0) = -g_2'(0)$.

Finally, by setting $P = \nabla^2_{\phi\theta} f(\phi^*, \theta^*)$ and $Q = \nabla^2_{\theta\theta} f(\phi^*, \theta^*)$, we get the results. □

E.2 Proof of Theorem 4

Since the Jacobian $A = \begin{bmatrix} 0 & -P \\ P^T & Q \end{bmatrix}$, by the definition of eigenvector equations we have

$$\begin{bmatrix} 0 & -P \\ P^T & Q \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \lambda \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

where y_1, y_2 and λ may be complex-valued. We can rewrite the above equations as follows:

$$-Py_2 = \lambda y_1 \tag{38}$$

$$P^T y_1 + Qy_2 = \lambda y_2 \tag{39}$$

Plugging Eq. (38) into Eq. (39) yields

$$\lambda^2 y_2 - \lambda Q y_2 + P^T P y_2 = 0 \tag{40}$$

Case 1. Consider $y_2 = 0$, then 1) if P has the full column rank, we have $y_1 = 0$ as well which violates the definition of eigenvectors; 2) if P does not have the full column rank, we have $\lambda = 0$.

Case 2. Consider $y_2 \neq 0$, we can multiply Eq. (40) by y_2^H (conjugate transpose of y_2) and then divide by $\|y_2\|^2$ in both sides, yielding

$$\lambda^2 - \frac{y_2^H Q y_2}{\|y_2\|^2}\lambda + \frac{y_2^H P^T P y_2}{\|y_2\|^2} = 0 \tag{41}$$

Let $a_1 = \frac{y_2^H Q y_2}{\|y_2\|^2}$ and $a_2 = \frac{y_2^H P^T P y_2}{\|y_2\|^2}$, by solving the equation $\lambda^2 - a_1\lambda + a_2 = 0$, we can get the results of (10). Next, we need to evaluate a_1 and a_2, respectively.

First note that $a_1 = \frac{y_2^H Q y_2}{\|y_2\|^2}$ is actually the *Rayleigh Quotient* of Q. Therefore, we consider a set of m orthonormal eigenvectors $\{x_{Q,i}\}_{i=1}^m$ corresponding to its m eigenvalues $\{\lambda_i(Q)\}_{i=1}^m$, and then there exists some set of m coefficients $\{b_i\}_{i=1}^n$, such that

$$y_2 = \sum_{i=1}^m b_i x_{Q,i}$$

where b_i may be complex-valued. Thus, we have

$$Qy_2 = \sum_{i=1}^m b_i \lambda_i(Q) x_{Q,i}$$

and

$$a_1 = \frac{\sum_{i=1}^m |b_i|^2 \lambda_i(Q)}{\sum_{i=1}^m |b_i|^2} = \sum_{i=1}^m \alpha_i \lambda_i(Q)$$

where we let $\alpha_i = \frac{|b_i|^2}{\sum_{i=1}^m |b_i|^2}$ for $i = 1, \cdots, m$, which satisfies $\alpha_i \geq 0$ and $\sum_{i=1}^m \alpha_i = 1$.

Similarly, as $a_2 = \frac{y_2^H P^T P y_2}{\|y_2\|^2}$ is a *Rayleigh Quotient* of $P^T P$, we have

$$a_2 = \sum_{i=1}^m \tilde{\alpha}_i \lambda_i(P^T P)$$

with $\tilde{\alpha}_i$ satisfying $\tilde{\alpha}_i \geq 0$ and $\sum_{i=1}^m \tilde{\alpha}_i = 1$.

Finally, if P does not have the full column rank, we can choose $y_2 \in \text{Null}(P)$ and $y_1 = 0$ such that $a_2 = 0$ and thus $\lambda = 0$ becomes a solution of Eq. (41). Therefore, the analysis of Case 1 is a special case of Case 2. □

E.3 Proof of Theorem 5

From Lemma 3, we know that for JARE, the corresponding regularization matrix is

$$\Gamma = \begin{bmatrix} I & -\gamma P \\ \gamma P^T & I \end{bmatrix}$$

Thus, the Jacobian becomes

$$\begin{aligned} A &= \Gamma \begin{bmatrix} 0 & -P \\ P^T & Q \end{bmatrix} \\ &= -\gamma \begin{bmatrix} PP^T & PQ \\ 0 & P^T P \end{bmatrix} + \begin{bmatrix} 0 & -P \\ P^T & Q \end{bmatrix} \end{aligned}$$

In the limit of $\gamma \to \infty$, we have

$$A = -\gamma \begin{bmatrix} PP^T & PQ \\ 0 & P^T P \end{bmatrix}$$

Its eigenvalues $\lambda(A)$ are solutions of $\det(\lambda I - A) = 0$. As a block upper triangular matrix, we have

$$\det(\lambda I - A) = \det(\lambda I + \gamma P^T P) \det(\lambda I + \gamma P P^T)$$

which means the eigenvalues of A satisfy

$$\lambda(A) = -\gamma \lambda(P^T P) \text{ and } \lambda(A) = -\gamma \lambda(P P^T)$$

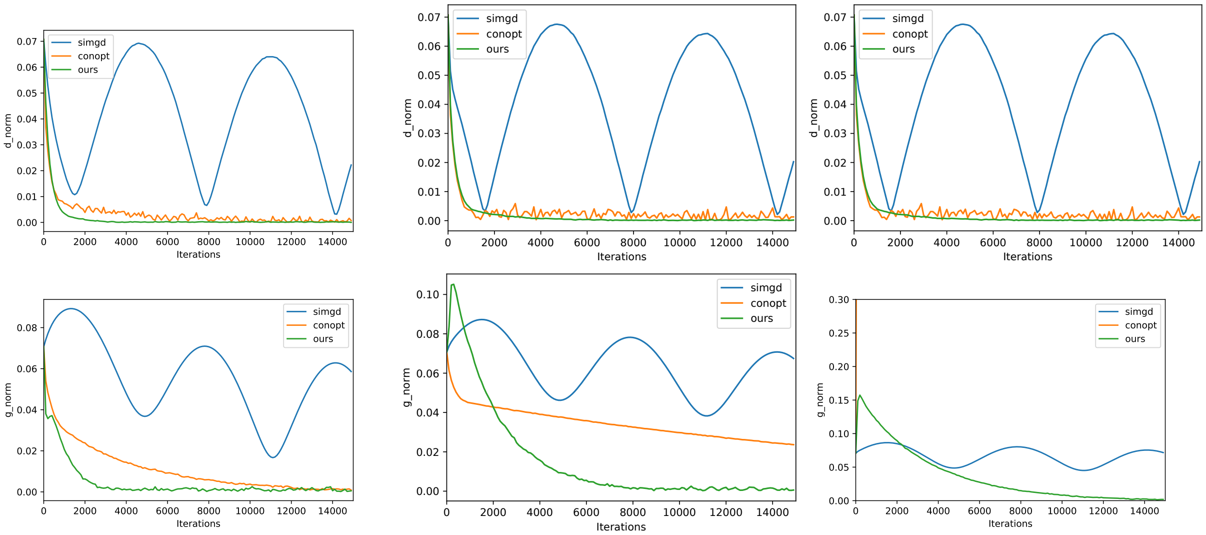
Also, since $P^T P$ and PP^T have the same set of eigenvalues, we have

$$\lambda(A) = -\gamma \lambda(P^T P)$$

as required. □

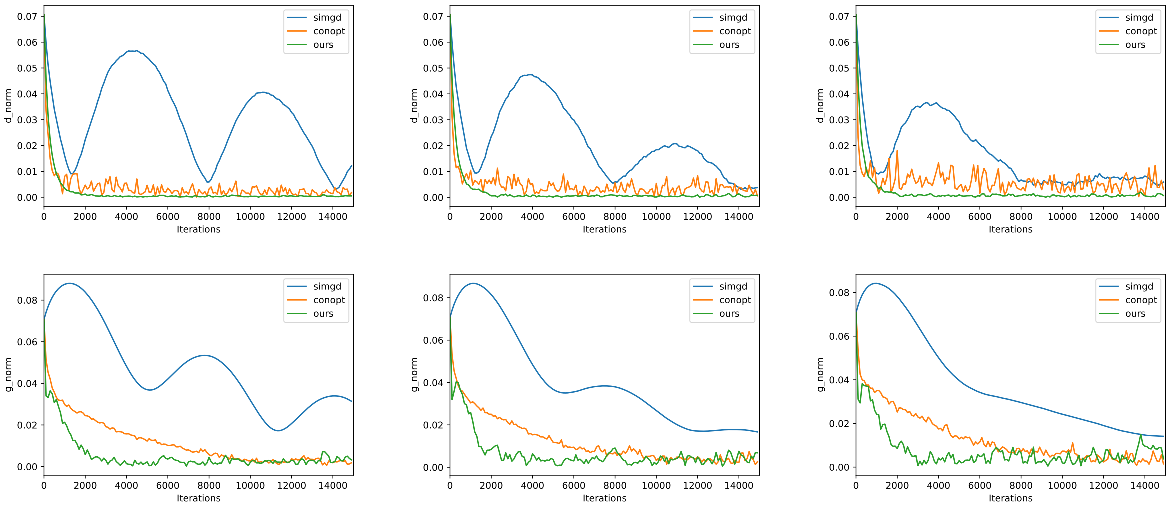
F More experimental results

F.1 More results on Isotropic Gaussian



(a) $\mu = 2$ (b) $\mu = 4$ (c) $\mu = 6$

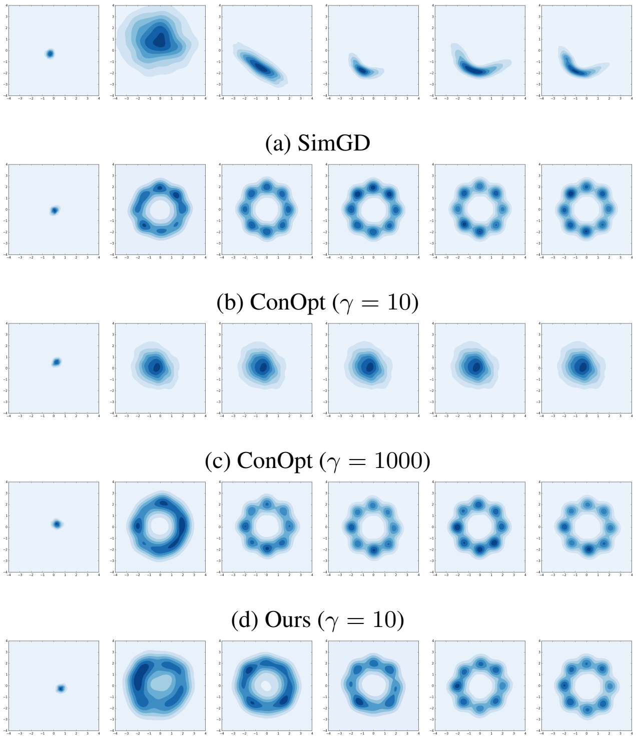
Figure 5: Training dynamics of SimGD, ConOpt and JARE for the discriminator (top row) and the generator (bottom row) with varying mean value μ where $\sigma = 0.2$. Note that as μ increases, the convergence rate for either SimGD or ConOpt becomes slower. When $\mu = 6$, the generator training curve for the ConOpt directly blow up.



(a) $\sigma = 0.5$ (b) $\sigma = 0.7$ (c) $\sigma = 0.9$

Figure 6: Training dynamics of SimGD, ConOpt and JARE for the discriminator (top row) and the generator (bottom row) with varying standard deviation σ where $\mu = 2$. Note that the damping effect in SimGD becomes stronger as the standard derivation σ increases.

F.2 More results on Mixture of Gaussians



(a) SimGD

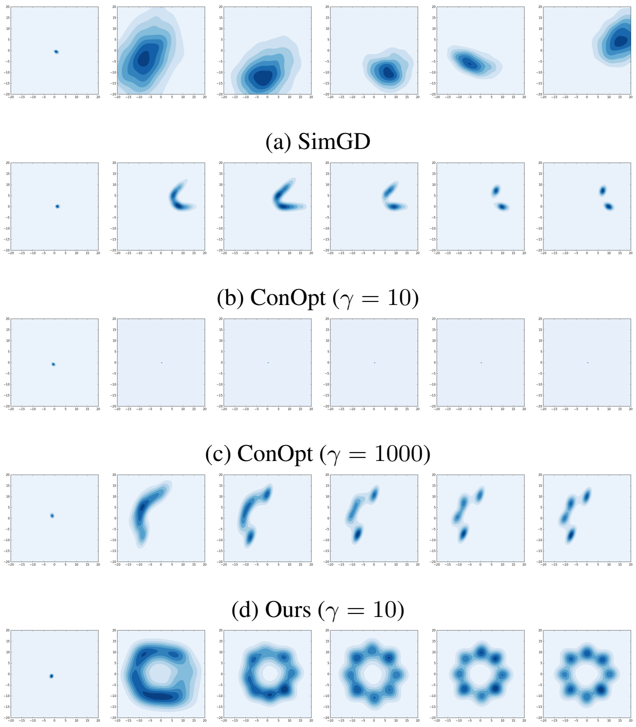
(b) ConOpt ($\gamma = 10$)

(c) ConOpt ($\gamma = 1000$)

(d) Ours ($\gamma = 10$)

(e) Ours ($\gamma = 1000$)

Figure 7: Comparison of SimGD (a), ConOpt (b,c) and Ours (d,e) on the mixture of Gaussians over iterations where $r = 2$. From left to right, each row consists of the results after 0, 2000, 4000, 6000, 8000 and 10000 iterations.



(a) SimGD

(b) ConOpt ($\gamma = 10$)

(c) ConOpt ($\gamma = 1000$)

(d) Ours ($\gamma = 10$)

(e) Ours ($\gamma = 1000$)

Figure 8: Comparison of SimGD (a), ConOpt (b,c) and Ours (d,e) on the mixture of Gaussians over iterations where $r = 10$. From left to right, each row consists of the results after 0, 2000, 4000, 6000, 8000 and 10000 iterations.

F.3 Network architectures

$z \in \mathbb{R}^{128} \sim \mathcal{N}(0, I)$
dense, $2 \times 2 \times M_f$
4×4, stride=2, deconv. M_f ReLU ResBlock M_f
4×4, stride=2, deconv. M_f ReLU ResBlock M_f
4×4, stride=2, deconv. M_f ReLU ResBlock M_f
4×4, stride=2, deconv. 3 tanh

(a) Generator

$x \in \mathbb{R}^{32 \times 32 \times 3}$
4×4, stride=2, conv. M_f ReLU ResBlock M_f
4×4, stride=2, conv. M_f ReLU ResBlock M_f
4×4, stride=2, conv. M_f ReLU ResBlock M_f
4×4, stride=2, conv. M_f ReLU ResBlock M_f
dense $\rightarrow 1$

(b) Discriminator

Table 2: ResNet architectures v1 for CIFAR-10 where M_f denotes the number of filters.

$z \in \mathbb{R}^{128} \sim \mathcal{N}(0, I)$
dense, $4 \times 4 \times M_f$
4×4, stride=2, deconv. M_f ReLU ResBlock M_f
4×4, stride=2, deconv. M_f ReLU ResBlock M_f
4×4, stride=2, deconv. M_f ReLU ResBlock M_f
3×3, stride=1, conv. 3 tanh

(a) Generator

$x \in \mathbb{R}^{32 \times 32 \times 3}$
4×4, stride=2, conv. M_f ReLU ResBlock M_f
4×4, stride=2, conv. M_f ReLU ResBlock M_f
4×4, stride=2, conv. M_f ReLU ResBlock M_f
4×4, stride=2, conv. M_f ReLU ResBlock M_f
dense $\rightarrow 1$

(b) Discriminator

Table 3: ResNet architectures v2 for CIFAR-10 where M_f denotes the number of filters.

$z \in \mathbb{R}^{128} \sim \mathcal{N}(0, I)$
dense, $4 \times 4 \times M_f$
4×4, stride=2, deconv. M_f ReLU ResBlock M_f
4×4, stride=2, deconv. M_f ReLU ResBlock M_f
4×4, stride=2, deconv. M_f ReLU ResBlock M_f
4×4, stride=2, deconv. M_f ReLU ResBlock M_f
3×3, stride=1, conv. 3 tanh

(a) Generator

$x \in \mathbb{R}^{64 \times 64 \times 3}$
4×4, stride=2, conv. M_f ReLU ResBlock M_f
4×4, stride=2, conv. M_f ReLU ResBlock M_f
4×4, stride=2, conv. M_f ReLU ResBlock M_f
4×4, stride=2, conv. M_f ReLU ResBlock M_f
4×4, stride=2, conv. M_f ReLU ResBlock M_f
dense $\rightarrow 1$

(b) Discriminator

Table 4: ResNet architectures for CelebA where M_f denotes the number of filters.

$z \in \mathbb{R}^{128} \sim \mathcal{N}(0, I)$
dense, $4 \times 4 \times M_f$
4×4, stride=2, deconv. M_f ReLU ResBlock M_f
4×4, stride=2, deconv. M_f ReLU ResBlock M_f
4×4, stride=2, deconv. M_f ReLU ResBlock M_f
4×4, stride=2, deconv. M_f ReLU ResBlock M_f
4×4, stride=2, deconv. M_f ReLU ResBlock M_f
3×3, stride=1, conv. 3 tanh

(a) Generator

$x \in \mathbb{R}^{128 \times 128 \times 3}$
4×4, stride=2, conv. M_f ReLU ResBlock M_f
4×4, stride=2, conv. M_f ReLU ResBlock M_f
4×4, stride=2, conv. M_f ReLU ResBlock M_f
4×4, stride=2, conv. M_f ReLU ResBlock M_f
4×4, stride=2, conv. M_f ReLU ResBlock M_f
4×4, stride=2, conv. M_f ReLU ResBlock M_f
dense $\rightarrow 1$

(b) Discriminator

Table 5: ResNet architectures for ImageNet where M_f denotes the number of filters.

F.4 Generated images on CIFAR-10 with four methods: GAN, SN-GAN, ConOpt and JARE.

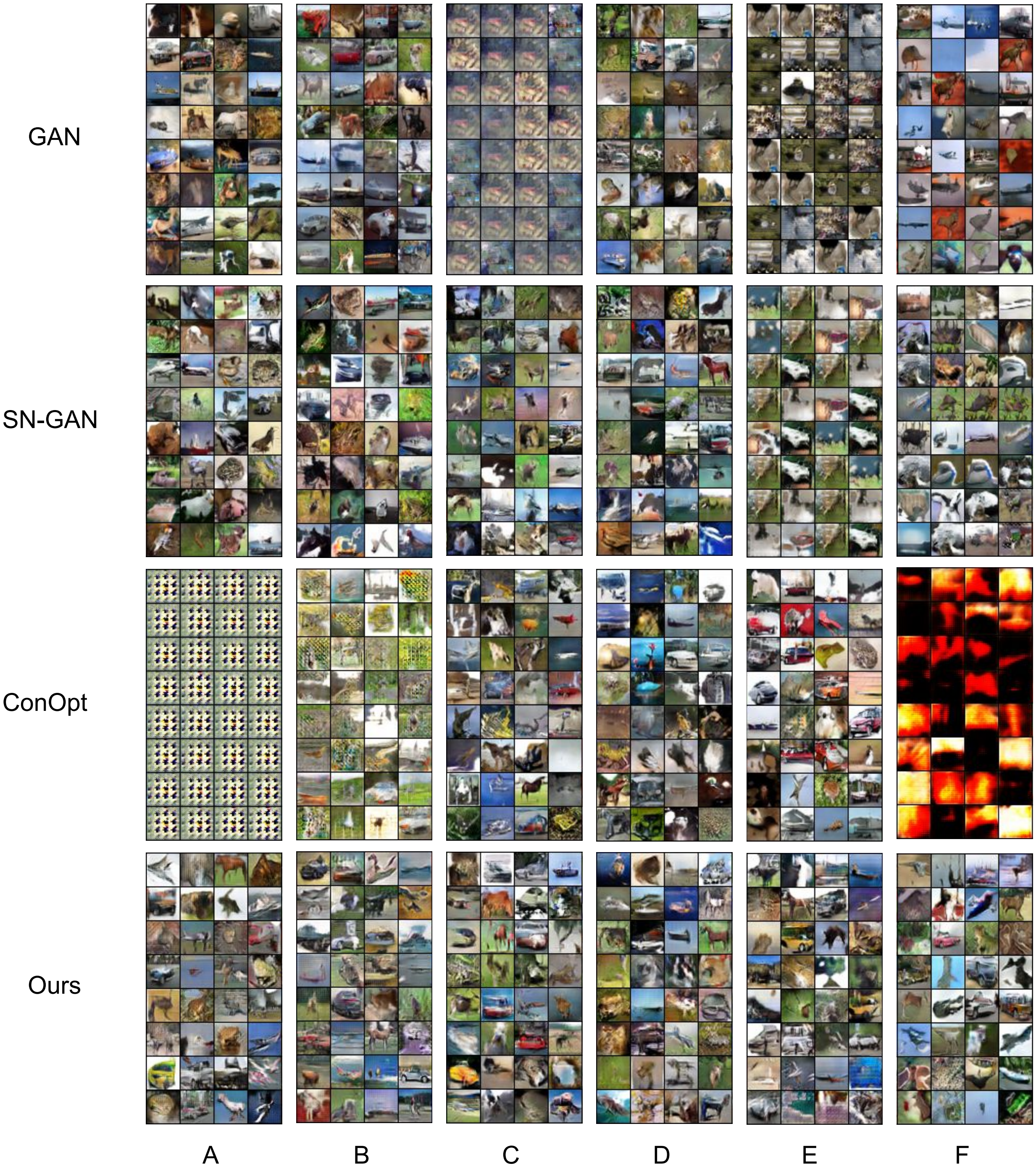
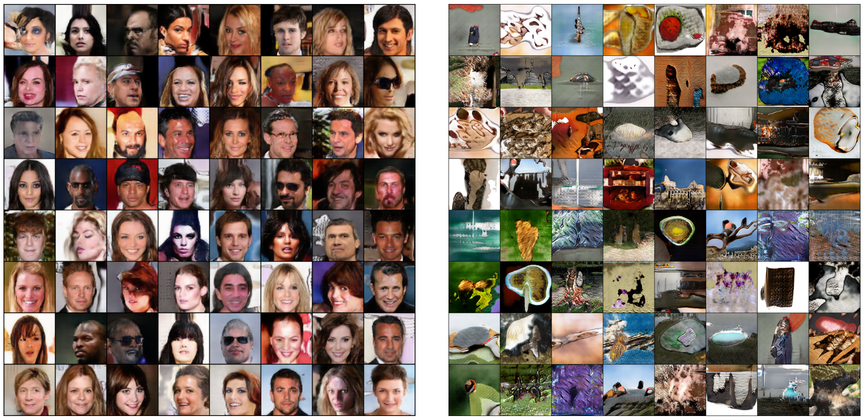


Figure 9: Generated images on CIFAR-10 with four training methods: standard GAN (or GAN), SN-GAN, ConOpt and JARE (Ours) in all the A-F settings. Best viewed in the electronic version by zooming in. We can see that only JARE is able to generate realistic images when training on CIFAR-10 across all six settings.



(a) Random samples for CelebA

(b) Random samples for ImageNet

Figure 10: Random samples generated by JAREs trained on CelebA and ImageNet, respectively, in an unsupervised manner. For CelebA, the sample size is 64×64, and for ImageNet, the sample size is 128×128.

F.5 More Rresults on CelebA and ImageNet.

In this experiment, we qualitatively evaluate the generated samples of JARE on the CelebA (with size of 64×64) (Liu et al., 2015) and ILSVRC2012 (ImageNet, with size of 128×128) (Russakovsky et al., 2015) datasets. Due to the limitation of our computational budgets, we do not apply large hyperparameter searches. Instead, we use a similar training setup as for the CIFAR-10 experiments, with slightly different network architectures. Please see Tables 4 and 5 in Appendix F.3 for details.

Figure 10 (a) and (b) show the randomly generated samples of JARE trained on CelebA and ImageNet, respectively. We can see that for CelebA, JARE can produce realistic and diverse celebrity faces with various backgrounds. For ImageNet, JARE can stabilize the training well while other training methods quickly collapse. While not completely realistic, it can generate visually convincing and diverse images from 1000 ImageNet classes in a completely unsupervised manner. The good results of JARE on CelebA and ImageNet demonstrate its ability of stabilizing the GAN training on more complex tasks.