ELeGANt: An Euler-Lagrange Analysis of Wasserstein Generative Adversarial Networks

Siddarth Asokan
Robert Bosch Center for Cyber-Physical Systems
Indian Institute of Science (IISc)
Bengaluru - 560012, India
siddartha@iisc.ac.in

Chandra Sekhar Seelamantula
Department of Electrical Engineering
Indian Institute of Science
Bengaluru - 560012, India
css@iisc.ac.in

Abstract

We consider Wasserstein generative adversarial networks (WGAN) with a gradient-norm penalty and analyze the underlying *functional* optimization problem within a variational setting. The optimal discriminator in this setting is the solution to a Poisson differential equation, and can be obtained in closed form without having to train a neural network. We illustrate this by employing a Fourier-series approximation to solve the Poisson differential equation. Experimental results based on synthesized low-dimensional Gaussian data demonstrate superior convergence behavior of the proposed approach in comparison with the baseline WGAN variants that employ weight-clipping, gradient or Lipschitz penalties on the discriminator. Further, within this setting, the optimal Lagrange multiplier can be computed in closed-form, and serves as a proxy for measuring GAN generator convergence. This work is an extended abstract, summarizing Asokan & Seelamantula (2023).

1 Introduction

The optimization of a generative adversarial network (GAN), originally proposed by Goodfellow et al. (2014), is a *min-max* game between two players — a generator (G) and a discriminator (D). The role of the generator is to create fake samples that mimic the ones coming from the training data distribution. The discriminator D is tasked with telling apart the real samples from the fake ones. The optimal G is the one that *outsmarts* D into confusing the fake samples for real. The generator G accepts high-dimensional noise $\boldsymbol{z} \sim p_\ell$ as input and generates fake samples $G(\boldsymbol{z}) \sim p_g$. The discriminator D accepts an input $\boldsymbol{x}$, which could come from either the data distribution p_d, or the generator distribution p_g, and outputs a value $D(\boldsymbol{x})$. Effectively, the generator must learn a mapping from the noise distribution to the data distribution, whereas the discriminator must learn the optimal two-class classifier. Over the past decade, numerous variants of GANs have been proposed with several successful applications.

GAN Losses: Almost all known GAN flavors minimize either a divergence metric (Goodfellow et al., 2014; Mao et al., 2017; Nowozin et al., 2016) or an integral probability metric. In the work, we consider integral probability metric (IPM) GANs, wherein the *discriminator* is a *real-valued critic* that differentiates between the generator and data distributions. The choice of the class of critics gives rise to variants such as the Wasserstein GAN (WGAN) (Arjovsky et al., 2017) with a Lipschitz-1 critic, the minimum-mean discrepancy GAN (MMD-GAN) (Li et al., 2017) where the critic is bounded by a ball in a reproducing-kernel Hilbert space, or the Fisher GAN (Mroueh & Sercu, 2017) in which the second-order moments of the critic are constrained to be bounded. Sobolev GANs (Mroueh et al., 2018) favor critics with a finite energy in the gradient. The constraints are enforced either by means of an adjustment of the network weights (Arjovsky et al., 2017; Roth et al., 2019; Wang & Liu, 2016), or through a suitable penalty incorporated into the loss function (Gulrajani et al., 2017; Roth et al., 2017; Mescheder et al., 2018). IPM based GANs have classically been analyzed

DLDE-III Workshop in the 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

within the framework of optimal transport (Sanjabi et al., 2018; Bousquet et al., 2017; Lei et al., 2019), while Unterthiner et al. (2018); Asokan & Seelamantula (2022) analyze the discriminator gradient-penalized GANs as a Coulomb potential function. Along a similar vein, Mroueh et al. (2018) established connections between the Sobolev GAN and the Fokker-Planck PDE. In this work, we show how the PDE connection can be leveraged to make the GAN optimization more insightful.

1.1 Our Contribution

This extended abstract summarizes the findings of Asokan & Seelamantula (2023). In the context of gradient-norm-penalized WGAN, leveraging the *Euler-Lagrange equation* from functional Calculus, we show that the optimal discriminator, given the generator, solves a Poisson PDE. The solution relates to n-D potential functions between the generator and data distributions. We propose to solve the discriminator PDE using a truncated Fourier-series model (and hence the name WGAN-FS), whose coefficients are obtained in closed form. This allows one to determine the optimal discriminator given the generator. We also show that the optimal value of the Lagrange multiplier can also be computed in closed form using a primal-dual approach, and tracking the optimal Lagrange multiplier becomes a viable alternative for measuring training convergence. Experimental validations on synthetic Gaussian datasets shows that training a GAN with the proposed Fourier-series discriminator outperforms baseline methods that consider a neural network discriminator. The source code is available online at `https://github.com/DarthSid95/ELF_GANs`

2 Gradient-regularized Wasserstein GANs

The WGAN minimizes *earth mover's distance* (EMD) between the generator and the target data distributions, p_g and p_d, respectively. Earth mover's distance is a special case of the Wasserstein distance between two distributions. Through Kantorovich-Rubinstein duality, the WGAN optimization is specified via the min-max problem:

$$\min_{p_g} \left\{ \max_{D} \left\{ \mathbb{E}_{\boldsymbol{x}\sim p_d}[D(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim p_g}[D(\boldsymbol{x})] \right\} \right\},$$

which is equivalent to the sequential minimization:

$$D^*(\boldsymbol{x}, p_g) = \arg \min_{D:\ \|D\|_L \leq 1} \mathcal{L}_D^{\mathrm{WGAN}}, \quad \text{where } \mathcal{L}_D^{\mathrm{WGAN}} = -\mathbb{E}_{\boldsymbol{x}\sim p_d}[D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[D(\boldsymbol{x})], \text{ and}$$
$$p_g^*(\boldsymbol{x}) = \arg \min_{p_g} \mathcal{L}_G^{\mathrm{WGAN}}, \quad \text{where } \mathcal{L}_G^{\mathrm{WGAN}} = \mathbb{E}_{\boldsymbol{x}\sim p_d}[D^*(\boldsymbol{x}, p_g)] - \mathbb{E}_{\boldsymbol{x}\sim p_g}[D^*(\boldsymbol{x}, p_g)]$$

where in turn, $\|D(\boldsymbol{x})\|_L \leq 1$ denotes the Lipschitz constraint on the discriminator and $D^*(\boldsymbol{x}, p_g)$ is the optimal discriminator for a given generator distribution p_g. The optimal discriminator D^* is the one that penalizes regions of the input space where p_g differs from p_d, while satisfying the Lipschitz constraint. The constraint is typically imposed by clipping the weights of the discriminator network. Subsequent work (Gulrajani et al., 2017; Kodali et al., 2017; Mescheder et al., 2018; Petzka et al., 2018; Mroueh et al., 2018; Terjék, 2020) replaced the Lipschitz constraint with a gradient penalty to avoid exploding gradients in a neural-network setting. For example, in WGAN-GP (Gulrajani et al., 2017), the gradients are evaluated over an interpolated distribution. Let $\mathcal{X}$ denote the convex hull that contains the supports of p_d and p_g. In this work, we consider the gradient-norm penalty (GNP):

$$\Omega_D : \int_{\mathcal{X}} \left(\|\nabla D(\boldsymbol{x})\|_2^2 - 1 \right) \mathrm{d}\boldsymbol{x}. \tag{1}$$

The proposed penalty can be viewed as a particular case of the penalty considered in the Sobolev GAN formulation. While WGAN-R$_d$ and WGAN-R$_g$ (Mescheder et al., 2018) enforce the penalty on the supports of p_d and p_g, respectively, the proposed WGAN-FS considers a uniform distribution on $\mathcal{X}$, resulting in a closed-form solution to the discriminator, given the generator. Incorporating Ω_D gives rise to the following regularized WGAN-FS discriminator cost:

$$\mathcal{L}_D = -\mathbb{E}_{\boldsymbol{x}\sim p_d}[D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[D(\boldsymbol{x})] + \lambda_d \int_{\mathcal{X}} \left(\|\nabla D(\boldsymbol{x})\|_2^2 - 1 \right) \mathrm{d}\boldsymbol{x}. \tag{2}$$

3 The Fourier-series-based WGAN-FS Formulation

Consider the WGAN-FS loss $\mathcal{L}_D$. The optimal discriminator corresponding to the loss given in Equation (2) is given by the following Theorem.

Theorem 3.1. ***Optimal WGAN-FS discriminator****: Consider the WGAN-FS discriminator loss subject to the gradient-norm penalty as given by Equation* (2)*. The optimizer of $\mathcal{L}_D$ solves Poisson's partial differential equation given by*

$$-\Delta D\left(\boldsymbol{x}\right) = \frac{p_d(\boldsymbol{x}) - p_g(\boldsymbol{x})}{2\lambda_d}, \tag{3}$$

where $\Delta = \nabla.\nabla = \left(\partial^2_{x_1} + \partial^2_{x_2} + \ldots + \partial^2_{x_n}\right)$ denotes the Laplacian operator.

Consider the integrand in Equation (2). Applying the Euler-Lagrange condition from the *Calculus of Variations* yields the optimality condition that the optimal discriminator solved a Poisson's partial differential equation (PDE) given in Equation (3). While we summarize the proof in Appendix C.1, additional details are given in (Asokan & Seelamantula, 2023). While kernel-based solutions to the above were explored by (Asokan & Seelamantula, 2022) in the context of RBF-CoulombGANs, in this work, we explore a Fourier-series-based solution.

Motivated by the Fourier series expansion to solve the heat equation in a metal (Fourier, 1807), based on the fact that they are eigenfunctions of the Laplace operator, we solve the discriminator PDE in Equation (3) using a Fourier-series expansions of p_d, p_g and $D(\boldsymbol{x})$:

$$p_d(\boldsymbol{x}) = \sum_{\boldsymbol{m}\in\mathbb{Z}^n} \alpha_{\boldsymbol{m}} e^{\mathrm{j}\langle \boldsymbol{\omega}_{\boldsymbol{m}}, \boldsymbol{x}\rangle},\ p_g(\boldsymbol{x}) = \sum_{\boldsymbol{m}\in\mathbb{Z}^n} \beta_{\boldsymbol{m}} e^{\mathrm{j}\langle \boldsymbol{\omega}_{\boldsymbol{m}}, \boldsymbol{x}\rangle},\ \text{and } D_{FS}(\boldsymbol{x}) = \frac{1}{\lambda_{FS}} \sum_{\boldsymbol{m}\in\mathbb{Z}^n} \gamma_{\boldsymbol{m}} e^{\mathrm{j}\langle \boldsymbol{\omega}_{\boldsymbol{m}}, \boldsymbol{x}\rangle}$$

with frequency harmonics $\boldsymbol{\omega}_{\boldsymbol{m}} = \omega_0\, \boldsymbol{m} = \omega_0[m_1, m_2, \ldots, m_n]^{\mathrm{T}},\ \boldsymbol{m} \in \mathbb{Z}^n - \{\boldsymbol{0}\}$. Substituting the Fourier-series expansions in (3) and comparing terms on both sides gives

$$\gamma_{\boldsymbol{m}} = \frac{1}{2}\left(\frac{\alpha_{\boldsymbol{m}} - \beta_{\boldsymbol{m}}}{\|\boldsymbol{\omega}_{\boldsymbol{m}}\|^2}\right), \quad \boldsymbol{m} \in \mathbb{Z}^n - \{\boldsymbol{0}\}. \tag{4}$$

The value of $\gamma_{\boldsymbol{0}}$ introduces a DC offset in $D_{FS}(\boldsymbol{x})$, and without loss of generality, we set $\gamma_{\boldsymbol{0}} = 0$. The Fourier coefficients of p_d and p_g are given by $\alpha_{\boldsymbol{m}} = \left(\frac{1}{T}\right)^n \varphi^*_{p_d}(\boldsymbol{\omega}_{\boldsymbol{m}})$ and $\beta_{\boldsymbol{m}} = \left(\frac{1}{T}\right)^n \varphi^*_{p_g}(\boldsymbol{\omega}_{\boldsymbol{m}})$, respectively, where φ^* represents the complex conjugate of the characteristic function of the corresponding distribution. The Fourier-series approximation yields the particular solution to the PDE. Including the family of homogeneous solutions $D_h(\boldsymbol{x}) = \langle \boldsymbol{a}, \boldsymbol{x}\rangle + a_{\text{constant}}$, the general solution becomes $D^*(\boldsymbol{x}) = D^*_{FS}(\boldsymbol{x}) + D_h(\boldsymbol{x})$. In computing $(\boldsymbol{a}, a_{\text{constant}})$, we have the following result:

Lemma 3.2. ***Optimal WGAN-FS generator****: Consider the optimization of the generator loss $\mathcal{L}_G = -\mathcal{L}_D$, with respect to p_g, where $D^*(\boldsymbol{x})$ is as given above. Then, the optimal solution is given by $p^*_g(\boldsymbol{x}) = p_d(\boldsymbol{x}),\ \forall\, \boldsymbol{x} \in \mathcal{X}$, and the solution is optimal for all finite real values of $(\boldsymbol{a}, a_{constant})$.*

The proof is given in Appendix C.2. Upon convergence of the GAN, $p^*_g(\boldsymbol{x}) = p_d(\boldsymbol{x})$, which implies $D^*_{opt}(\boldsymbol{x}) = D^*_h(\boldsymbol{x}) = 0$. Therefore, without loss of optimality, we set $(\boldsymbol{a}, a_{\text{constant}}) = (\boldsymbol{0}, 0)$.

For $n \geq 4$, in order to reduce the number of terms in the Fourier summation, we consider two truncation frequencies, M_{low} and M_{high}. We deterministically include all low-frequency components along each dimension to M_{low}, while uniformly sampling coefficients between M_{low} and M_{high} (together denoted by the set $\mathcal{M}$). Further, from an implementation perspective, to avoid complex arithmetic, we use a trigonometric Fourier-series expansion, and the resulting Fourier-series discriminator is:

$$D^*_{FS}(\boldsymbol{x}) \approx \frac{1}{\lambda^*_{FS}}\left(\frac{\gamma_{\boldsymbol{0}}}{2} + \sum_{\boldsymbol{m}\in\mathcal{M}} \gamma^r_{\boldsymbol{m}} \cos(\omega_o\langle \boldsymbol{m}, \boldsymbol{x}\rangle) + \sum_{\boldsymbol{m}\in\mathcal{M}} \gamma^i_{\boldsymbol{m}} \sin(\omega_o\langle \boldsymbol{m}, \boldsymbol{x}\rangle)\right), \tag{5}$$

where $\gamma^r_{\boldsymbol{m}}$ and $\gamma^i_{\boldsymbol{m}}$ are the real and imaginary part of $\gamma_{\boldsymbol{m}}$, respectively. Enforcing the gradient-norm penalty Ω_D on (5) results in a bound on λ^*_{FS}. The worst-case value of λ^*_{FS} satisfies:

$$\lambda^*_{FS} \approx \sqrt{(2|\mathcal{M}| + 1)\left(\sum_{\boldsymbol{m}\in\mathcal{M}} (\tau^i_{\boldsymbol{m}} + \tau^r_{\boldsymbol{m}}) + \frac{1}{N}\sum_{k=1}^{N}\sum_{\boldsymbol{m}\in\mathcal{M}} (\tau^i_{\boldsymbol{m}} - \tau^r_{\boldsymbol{m}})\cos(2\omega_o\langle \boldsymbol{m}, \boldsymbol{x}_k\rangle)\right)},$$

where $\tau^r_{\boldsymbol{m}} = \frac{1}{2}(\gamma^r_{\boldsymbol{m}})^2\omega_o^2\|\boldsymbol{m}\|^2$, and $\tau^i_{\boldsymbol{m}} = \frac{1}{2}(\gamma^i_{\boldsymbol{m}})^2\omega_o^2\|\boldsymbol{m}\|^2$, and the samples $\boldsymbol{x}_k$ are drawn from the uniform mixture of p_d and p_g. Based on the approximation- and truncation-error analysis presented by Asokan & Seelamantula (2023), we set $M_{low} = 2$, $M_{high} = 10$, and $L = 10^3$.

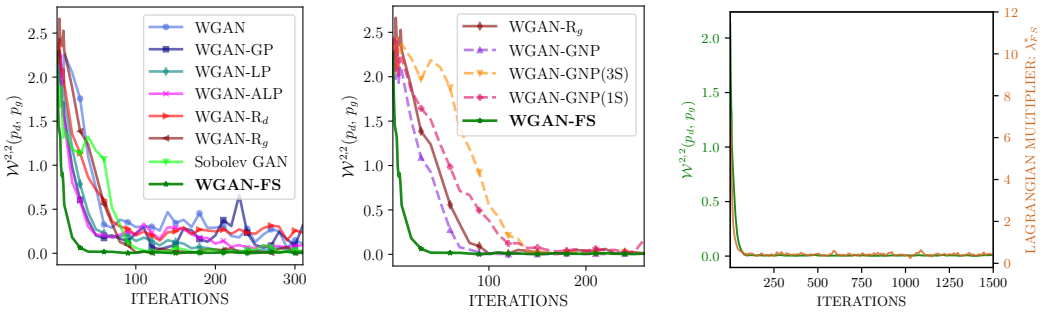

Figure 1: (Color online) Experiments on 2-D Gaussian data: (a) & (b) Wasserstein-2 distance ($\mathcal{W}^{2,2}(p_d, p_g)$) between WGAN-FS and (a) baseline WGAN variants, (b) trainable variants of the proposed WGAN-FS. The closed-form Fourier-series approach to enforcing the gradient-norm penalty converges an order faster than the baselines and trainable variants of the same loss. (c) Convergence of the optimal Lagrange multiplier λ^*_{FS} alongside Wasserstein-2 distance between p_d and p_g ($\mathcal{W}^{2,2}(p_d, p_g)$). When the model appears converge in the sense of $\mathcal{W}^{2,2}(p_d, p_g)$, it is a measure of second-order statistics, while for λ^*_{FS}, the distributions converge in the L_2 sense (the Fourier representation of p_g converging to that of p_d).

3.1 Illustration Using Synthetic 2-D Data

As part of this extended abstract, we demonstrate performance on 2-D Gaussian learning problems. An in depth experimental validation, on latent-space modeling with Wasserstein autoencoders, is presented in Journal version (Asokan & Seelamantula, 2023).

We conduct experiments on 2-D Gaussian and 8-component Gaussian mixture models (GMM). We draw Gaussian data from $\mathcal{N}(0.75\mathbf{1}_2, 0.1\mathbb{I}_2)$, where $\mathbf{1}_2$ denotes a 2-D vector with both entries equal to 1, and $\mathbb{I}_2$ denotes the 2×2 identity matrix. The noise that is input to the generator is drawn from a Gaussian $\mathcal{N}(\mathbf{0}_2, \mathbb{I}_2)$. The baselines are discussed in Appendices B and D.1, while training specifications are provided in Appendix D.1.

Figures 1(a) and (b) show the Wasserstein-2 distance $\mathcal{W}^{2,2}(p_d, p_g)$ between the generator and true data distributions as a function of the iterations for the WGAN and WGAN-FS flavors under consideration, respectively, for 2-D Gaussian data. The Wasserstein-2 distance decays much faster in the case of WGAN-FS compared with the baseline variants. We observe that replacing the baseline gradient-norm penalty with that of WGAN-FS (denoted by WGAN-GNP) results in a performance on par with the best-case baseline. Similarly, training a single-layer discriminator with a sinusoidal activation function (WGAN-GNP (1S)) to approximately learn the Fourier coefficients results in poorer performance compared with WGAN-FS, as the suboptimal coefficients cannot represent the distributions p_d or p_g accurately. Figure 1(c) shows λ^*_{FS} and the Wasserstein-2 distance ($\mathcal{W}^{2,2}$) between p_d and p_g as a function of iterations. While $\mathcal{W}^{2,2}(p_d, p_g)$ measures second-order statistics between p_d and p_g, λ^*_{FS} measures the coefficient-wise convergence between the Fourier-series of p_d and p_g, which indirectly measures the L_2 error between the generator and target densities.

4 Discussions and Conclusion

In this work, we analyzed the Wasserstein GAN subjected to a novel variant of the gradient-norm penalty, leveraging results from functional Calculus. Within this framework, the optimal discriminator was shown to be the solution to a second-order partial differential equation (PDE). The PDE connection for the optimal discriminator provides a novel viewpoint for GAN optimization. By employing a Fourier-series approximation, we showed that a *single-shot* solution can be obtained for the discriminator, given the generator. The solution relies on the estimates of the characteristic functions of the data and generator distributions. The superior performance of this novel approach was demonstrated in low-dimensional multivariate Gaussian settings, while experiments on latent-space image data, and the variational analysis of various f-GAN (Nowozin et al., 2016) variants are presented in the Journal version. The choice of Fourier bases was motivated by Laplace operator and employing alternative bases representations is a promising direction for future research.

Acknowledgements

This work is supported by the Microsoft Research Ph.D. Fellowship, the Qualcomm Innovation Fellowship 2023, the Robert Bosch Center for Cyber-Physical Systems Ph.D. Fellowship, and the Science and Engineering Research Board (SERB) Core Research Grant (CoRE), Department of Science and Technology.

References

Abadi, M. et al. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint, arXiv:1603.04467*, Mar. 2016. URL `https://arxiv.org/abs/1603.04467`.

Adler, J. and Lunz, S. Banach Wasserstein GAN. In *Advances in Neural Information Processing Systems 31*, pp. 6754–6763. 2018.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 214–223, 2017.

Asokan, S. and Seelamantula, C. S. Bridging the gap between Coulomb GAN and gradient-regularized WGAN. In *Proceedings on "The Symbiosis of Deep Learning and Differential Equations - II" at NeurIPS Workshops*, 2022.

Asokan, S. and Seelamantula, C. S. Euler-Lagrange analysis of generative adversarial networks. *Journal of Machine Learning Research (JMLR)*, pp. 1–100, 2023.

Bousquet, O., Gelly, S., Tolstikhin, I., Simon-Gabriel, C. J., and Schoelkopf, B. From optimal transport to generative modeling: the VEGAN cookbook. *arXiv preprint, arXiv:1705.07642*, May 2017. URL `https://arxiv.org/abs/1705.07642`.

Evans, L. C. *Partial Differential Equations*. American Mathematical Society, 2010.

Ferguson, J. A brief survey of the history of the calculus of variations and its applications. *arXiv preprint, arXiv:math/0402357*, Feb. 2004. URL `https://arxiv.org/abs/math/0402357`.

Flamary, R. et al. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78): 1–8, 2021. URL `http://jmlr.org/papers/v22/20-451.html`.

Fourier, J. B. J. Mémoire sur la propagation de la chaleur dans les corps solides. *Présenté le 21 décembre 1807 à l'Institut national - Nouveau Bulletin des sciences par la Sociétéphilomatique de Paris*, pp. 215–221, 1807.

Gel'fand, I. M. and Fomin, S. V. *Calculus of Variations*. Prentice-Hall, 1964.

Genevay, A., Peyre, G., and Cuturi, M. Learning generative models with Sinkhorn divergences. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84, pp. 1608–1617. PMLR, 2018.

Goldstine, H. H. *A History of the Calculus of Variations from the 17th Through the 19th Century*. Springer, New York, 1980.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. 2014.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems 30*, pp. 5767–5777. 2017.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

Kodali, N., Abernethy, J. D., Hays, J., and Kira, Z. On convergence and stability of GANs. *arXiv preprint, arXiv:1705.07215*, May 2017. URL `http://arxiv.org/abs/1705.07215`.

Lei, N., Guo, Y., An, D., Qi, X., Luo, Z., Yau, S. T., and Gu, X. Mode collapse and regularity of optimal transportation maps. *arXiv preprints, arXiv:1902.02934*, Feb. 2019. URL https://arxiv.org/abs/1902.02934.

Li, C. L., Chang, W. C., Cheng, Y., Yang, Y., and Poczos, B. MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems 30*, pp. 2203–2213. 2017.

Mao, X., Li, Q., Xie, H., Lau, R. Y. K., Wang, Z., and Smolley, S. P. Least squares generative adversarial networks. In *Proceedings of International Conference on Computer Vision*, 2017.

Mescheder, L., Geiger, A., and Nowozin, S. Which training methods for GANs do actually converge? In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3481–3490, Stockholmsmassan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Mroueh, Y. and Sercu, T. Fisher GAN. In *Advances in Neural Information Processing Systems 30*, pp. 2513–2523. 2017. URL http://arxiv.org/abs/1705.09675.

Mroueh, Y., Li, C., Sercu, T., Raj, A., and Cheng, Y. Sobolev GAN. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

Nowozin, S., Cseke, B., and Tomioka, R. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems 29*, pp. 271–279. 2016.

Petzka, H., Fischer, A., and Lukovnikov, D. On the regularization of Wasserstein GANs. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

Roth, K., Lucchi, A., Nowozin, S., and Hofmann, T. Stabilizing training of generative adversarial networks through regularization. In *Advances in Neural Information Processing Systems 30*, pp. 2015–2025, 2017.

Roth, K., Kilcher, Y., and Hofmann, T. Adversarial training generalizes data-dependent spectral norm regularization. *arXiv preprint, arXiv:1906.01527*, June 2019. URL https://arxiv.org/abs/1906.01527.

Sanjabi, M., Ba, J., Razaviyayn, M., and Lee, J. D. On the convergence and robustness of training GANs with regularized optimal transport. In *Advances in Neural Information Processing Systems 31*, pp. 7091–7101. 2018.

Sitzmann, V., Martel, J., Bergman, A., Lindell, D., and Wetzstein, G. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7462–7473, 2020.

Terjék, D. Adversarial Lipschitz regularization. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.

Tolstikhin, I. O., Bousquet, O., Gelly, S., and Schölkopf, B. Wasserstein auto-encoders. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

Unterthiner, T., Nessler, B., Seward, C., Klambauer, G., Heusel, M., Ramsauer, H., and Hochreiter, S. Coulomb GANs: Provably optimal Nash equilibria via potential fields. In *Proceedings of the 6th International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SkVqXOxCb.

Wang, D. and Liu, Q. Learning to draw samples: with application to amortized MLE for generative adversarial learning. *arXiv preprint, arXiv:1611.01722*, Nov. 2016. URL https://arxiv.org/abs/1611.01722.

Appendix

Table of Contents

A Mathematical Preliminaries

The cornerstone of our analysis is the Euler-Lagrange (EL) framework, which is at the heart of *Calculus of Variations* (Gel'fand & Fomin, 1964). The EL conditions are of fundamental importance in solving several problems in physics (Goldstine, 1980; Ferguson, 2004).

Consider the functional optimization of a cost $\mathcal{L}$ defined as

$$\mathcal{L}\left(y(x), y'(x)\right) = \int_a^b \mathcal{F}\left(x, y(x), y'(x)\right)\, \mathrm{d}x, \tag{6}$$

with respect to $y(x)$, $x \in [a, b]$, which is assumed to be continuously differentiable or at least continuous with a piecewise-smooth derivative $y'(x)$, with finite Dirichlet boundary conditions. Let $y^*(x)$ denote the optimizer of $\mathcal{L}$. The first variation of $\mathcal{L}$ at the optimum y^*, is defined as the Gateaux derivative $\delta\mathcal{L}(y^*, \eta) = \left.\dfrac{\partial \mathcal{L}_\epsilon(y^*)}{\partial \epsilon}\right|_{\epsilon=0}$, where

$$\mathcal{L}_\epsilon(y^*) = \mathcal{L}\left(y^*(x) + \epsilon\,\eta(x), y^{*'}(x) + \epsilon\,\eta'(x)\right) = \int_a^b \mathcal{F}\left(x, y^*(x) + \epsilon\,\eta(x), y^{*'}(x) + \epsilon\,\eta'(x)\right)\, \mathrm{d}x,$$

where, in turn, $\eta(x)$ is a family of compactly supported, infinitely differentiable functions that are identically zero at the boundaries $x = a$ and $x = b$. Setting the first variation to zero and invoking the fundamental lemma of Calculus of Variations gives rise to the Euler-Lagrange condition. The *fundamental lemma of Calculus of Variations* states that if a function $f(x)$ satisfies the condition

$$\int_a^b f(x)\,\eta(x)\,\mathrm{d}x = 0$$

for all compactly supported, infinitely differentiable functions $\eta(x)$, then f must be identically zero almost everywhere in $[a, b]$.

The Euler-Lagrange condition that the optimizer $y^*(x)$ must satisfy is given as follows:

$$\left.\frac{\partial \mathcal{F}}{\partial y} - \frac{\partial}{\partial x}\left(\frac{\partial \mathcal{F}}{\partial y'}\right)\right|_{y=y^*(x)} = 0. \tag{7}$$

In the special case where the cost $\mathcal{L}$ does not involve the derivative of y, the EL condition reduces to the degenerate version:

$$\left.\frac{\partial \mathcal{F}}{\partial y}\right|_{y=y^*(x)} = 0,$$

which simply corresponds to a point-wise optimization of y over $x \in [a, b]$.

In the multivariate case, that is, $\boldsymbol{x} \in \mathbb{R}^n$, the cost is of the type

$$\mathcal{L}\left(y\left(\boldsymbol{x}\right), \{y_i'\}_{i=1}^n\right) = \int\limits_{\mathcal{X} \subseteq \mathbb{R}^n} \mathcal{F}\left(\boldsymbol{x}, y, \{y_i'\}_{i=1}^n\right) \, \mathrm{d}\boldsymbol{x},$$

where $\mathcal{X}$ is the domain of integration and y_i' denotes the partial derivative of $y(\boldsymbol{x})$ w.r.t. the i^{th} entry of $\boldsymbol{x}$, that is, x_i. The corresponding EL condition is

$$\left.\frac{\partial \mathcal{F}}{\partial y} - \sum_{i=1}^{N} \left[\frac{\partial}{\partial x_i}\left(\frac{\partial \mathcal{F}}{\partial y_i'}\right)\right]\right|_{y=y^*(\boldsymbol{x})} = 0. \tag{8}$$

The EL condition is a first-order condition and enforcing it yields the optimum. Whether the optimum corresponds to a minimizer or maximizer of the cost must be checked by invoking the second-order condition, more specifically the Legendre-Clebsch necessary condition for a minimizer. In the 1-D case, the condition is given by $\dfrac{\partial^2 \mathcal{F}}{\partial y'^2} \geq 0$. In the multivariate setting, this condition translates to the positive-semi-definiteness (p.s.d.) of the Hessian matrix $\mathbb{H}$ of the Hamiltonian $\mathcal{H}$, computed with respect to $\{y_i'(\boldsymbol{x})\}_{i=1}^n$ and evaluated at $y(\boldsymbol{x}) = y^*(\boldsymbol{x})$: $\left.\mathbb{H}_{y,\mathcal{H}}\right|_{y=y^*} \succ 0$, where $\succ$ denotes the p.s.d. property. The Hamiltonian is given by

$$\mathcal{H} = \sum_{i=1}^{n} \left(y_i' \frac{\partial \mathcal{F}}{\partial y_i'}\right) - \mathcal{F},$$

and the entries of the Hessian are given by

$$[\mathbb{H}_{y,\mathcal{H}}]_{i,j} = \frac{\partial^2 \mathcal{H}}{\partial y_i' \partial y_j'}.$$

Together, these results are leveraged to derive the optimal discriminator in WGAN-FS.

B An Overview of Wasserstein GANs

The WGAN minimizes *earth mover's distance* (EMD) between the generator and the target data distributions, p_g and p_d, respectively. Earth mover's distance is a special case of the Wasserstein distance between two distributions. Through Kantorovich-Rubinstein duality, the WGAN optimization is specified via the min-max problem:

$$\min_{p_g} \left\{ \max_{D} \left\{ \mathbb{E}_{\boldsymbol{x} \sim p_d}[D(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x} \sim p_g}[D(\boldsymbol{x})] \right\} \right\},$$

which is equivalent to the sequential minimization:

$$D^*(\boldsymbol{x}, p_g) = \arg \min_{D:\ \|D\|_L \leq 1} \mathcal{L}_D^{\text{WGAN}}, \quad \text{where } \mathcal{L}_D^{\text{WGAN}} = -\mathbb{E}_{\boldsymbol{x} \sim p_d}[D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x} \sim p_g}[D(\boldsymbol{x})], \text{ and}$$

$$p_g^*(\boldsymbol{x}) = \arg \min_{p_g} \mathcal{L}_G^{\text{WGAN}}, \quad \text{where } \mathcal{L}_G^{\text{WGAN}} = \mathbb{E}_{\boldsymbol{x} \sim p_d}[D^*(\boldsymbol{x}, p_g)] - \mathbb{E}_{\boldsymbol{x} \sim p_g}[D^*(\boldsymbol{x}, p_g)]$$

where in turn, $\|D(\boldsymbol{x})\|_L \leq 1$ denotes the Lipschitz constraint on the discriminator and $D^*(\boldsymbol{x}, p_g)$ is the optimal discriminator for a given generator distribution p_g. The optimal discriminator D^* is the one that penalizes regions of the input space where p_g differs from p_d, while satisfying the Lipschitz constraint. The constraint is typically imposed by clipping the weights of the discriminator network.

Table 1: Discriminator loss functions corresponding to various WGAN variants considered in the literature alongside the proposed WGAN with gradient-norm penalty (WGAN-FS). The key difference lies in how the Lipschitz penalty is enforced on the discriminator. While the vanilla WGAN clips the discriminator network weights, the other WGAN flavors, including ours, consider gradient-based regularization.

WGAN flavor	Discriminator loss
WGAN	$\mathcal{L}_D^{\text{WGAN}} = -\,\mathbb{E}_{\boldsymbol{x}\sim p_d}[D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[D(\boldsymbol{x})]$
WGAN-GP	$\mathcal{L}_D^{\text{WGAN}} + \lambda\,\mathbb{E}_{\boldsymbol{x}\sim \alpha p_g + (1-\alpha)p_d}\left[(\|\nabla D(\boldsymbol{x})\|_2 - 1)^2\right];\ 0 \le \alpha \le 1$
WGAN-R$_d$R$_g$	$\mathcal{L}_D^{\text{WGAN}} + \frac{\lambda_1}{2}\,\mathbb{E}_{\boldsymbol{x}\sim p_d}\left[\|\nabla D(\boldsymbol{x})\|_2^2\right] + \frac{\lambda_2}{2}\,\mathbb{E}_{\boldsymbol{x}\sim p_g}\left[\|\nabla D(\boldsymbol{x})\|_2^2\right]$
Sobolev GAN	$\mathcal{L}_D^{\text{WGAN}} + \lambda\,\mathbb{E}_{\boldsymbol{x}\sim \nu_p(\boldsymbol{x})}\left[\|\nabla D(\boldsymbol{x})\|_2^2\right]$, where $\nu_p(\boldsymbol{x}) \ge 0;\ \int_{\mathcal{X}} \nu_p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$
WGAN-LP	$\mathcal{L}_D^{\text{WGAN}} + \lambda\,\mathbb{E}_{\boldsymbol{x}\sim \alpha p_g + (1-\alpha)p_d}\left[(\max(\|\nabla D(\boldsymbol{x})\|_2 - 1, 0))^2\right];\ 0 \le \alpha \le 1$
WGAN-ALP	$\mathcal{L}_D^{\text{WGAN}} + \lambda\,\mathbb{E}_{\boldsymbol{x}\sim p_d}\left[\left(\max\left(\frac{D(\boldsymbol{x}) - D(\boldsymbol{x}+\boldsymbol{r}_{adv})}{\|\boldsymbol{r}_{adv}\|_2} - 1, 0\right)\right)^2\right]$, where $\boldsymbol{r}_{adv} = \max_{\boldsymbol{r}:\|\boldsymbol{r}\|_2>0}\left\{\frac{D(\boldsymbol{x}) - D(\boldsymbol{x}+\boldsymbol{r})}{\|\boldsymbol{r}\|_2}\right\}$
WGAN-FS (Proposed)	$\mathcal{L}_D^{\text{WGAN}} + \lambda_d \int_{\boldsymbol{x}\in\mathcal{X}} \left(\|\nabla D(\boldsymbol{x})\|_2^2 - 1\right)\mathrm{d}\boldsymbol{x}$

An alternative to weight-clipping is spectral normalization of the weights (Roth et al., 2019). Subsequent works (Gulrajani et al., 2017; Petzka et al., 2018; Terjék, 2020) replaced the Lipschitz constraint with a gradient penalty to avoid exploding gradients in a neural-network setting. For example, Gulrajani et al. (2017) replaced the Lipschitz-1 penalty with the gradient penalty (WGAN-GP): $(\|\nabla D(\boldsymbol{x})\|_2 - 1)^2 = 0$. It is well-known that a function whose gradient has a bounded norm satisfies the Lipschitz constraint (Adler & Lunz, 2018).

Table 1 lists a few important gradient-based regularizers proposed in the WGAN literature, which are considered in this paper. The original WGAN-GP empirically evaluated the discriminator gradient on samples drawn from the interpolated distribution $\alpha p_g + (1-\alpha)p_d$, $0 \le \alpha \le 1$, and penalizes values far away from 1 in the norm-squared sense. Petzka et al. (2018) incorporated a *one-sided hinge-like* penalty in the WGAN-LP formulation (LP stands for Lipschitz penalty). The gradient magnitude is upper-bounded by 1, by penalizing the discriminator only when the gradient magnitude exceeds 1. The gradients were evaluated empirically on an interpolated distribution as in the case of WGAN-GP. In the adversarial Lipschitz regularization proposed in WGAN-ALP (Terjék, 2020), for a sample drawn from either the data or generator distributions, the regularizer was evaluated along the *adversarial* penalty direction $\boldsymbol{r}_{adv}$ — the one along which the Lipschitz constraint is maximally violated.

Mroueh et al. (2018) considered a gradient-norm penalty in the Sobolev GAN formulation, where they bounded the energy in the gradient of the discriminator, evaluated with respect to a base measure $\nu_p(\boldsymbol{x})$. From an implementation standpoint, they considered two base measures: (a) The midpoint distribution $\nu_p(\boldsymbol{x}) = \frac{p_d + p_g}{2}$, which is a special case of the WGAN-GP penalty (Gulrajani et al., 2017); and (b) A noise-convolved version of p_d, also considered in DRAGAN (Kodali et al., 2017). Mescheder et al. (2018) employed gradient penalties evaluated independently over real data (WGAN-R$_d$), over the generated data (WGAN-R$_g$), or a weighted combination of both (WGAN-R$_d$R$_g$) which can be seen as special cases of the Sobolev GAN penalty. Subsequent works extended the Wasserstein-1 distance based GAN to general L_p-norm spaces (Adler & Lunz, 2018) or propose solving the primal problem through Sinkhorn fixed-point iterations (Genevay et al., 2018).

C Optimality of WGAN-FS

In this appendix, we present the proofs for the optimal discriminator, generator, and Fourier-series-based Lagrange multiplier in WGAN-FS.

C.1 Optimal WGAN-FS Discriminator

Consider the n-dimensional WGAN-FS scenario, the discriminator loss takes the form:

$$\begin{aligned}\mathcal{L}_D &= -\mathbb{E}_{\boldsymbol{x}\sim p_d}[D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[D(\boldsymbol{x})] + \lambda_d \int_{\mathcal{X}} \left(\|\nabla D(\boldsymbol{x})\|_2^2 - 1\right)\, \mathrm{d}\boldsymbol{x} \\ &= \int_{\mathcal{X}} \left(D(\boldsymbol{x})\left(p_g(\boldsymbol{x}) - p_d(\boldsymbol{x})\right) + \lambda_d \left(\|\nabla D(\boldsymbol{x})\|_2^2 - 1\right)\right)\, \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{9}$$

To determine the optimal discriminator corresponding to the loss given in Equation (9), Consider the integrand: $D(\boldsymbol{x})(p_g(\boldsymbol{x}) - p_d(\boldsymbol{x})) + \lambda_d \|\nabla D(\boldsymbol{x})\|_2^2$. Applying the Euler-Lagrange condition from Equation (8) for obtaining the optimum results in Poisson's partial differential equation (PDE) given in Equation (3).

A closed-form solution to Poisson's equation is obtained similar to Coulomb GAN (Unterthiner et al., 2018) and RBF-Coulomb GAN formulations (Asokan & Seelamantula, 2022), by solving the n-D inhomogeneous differential equation $-\Delta D(\boldsymbol{x}) = \delta(\boldsymbol{x})$. In polar coordinates, this yields the fundamental solution $\phi(\boldsymbol{x})$ given by (Evans, 2010):

$$\phi(\boldsymbol{x}) = \begin{cases} -\dfrac{1}{2\pi}\ln(\|\boldsymbol{x}\|), & \text{for } n = 2, \text{ and} \\ \dfrac{1}{n(n-2)\mathfrak{v}(n)}\dfrac{1}{\|\boldsymbol{x}\|^{n-2}}, & \text{for } n \geq 3, \end{cases} \tag{10}$$

where $\|\boldsymbol{x}\| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$ and $\mathfrak{v}(n)$ is the volume of the unit sphere in $\mathbb{R}^n$ given by $\mathfrak{v}(n) = \pi^{\frac{n}{2}}\left(\Gamma\left(\frac{n}{2}+1\right)\right)^{-1}$, with $\Gamma(n)$ denoting the gamma function. The solution to Poisson's equation $-\Delta D(\boldsymbol{x}) = \frac{p_d(\boldsymbol{x}) - p_g(\boldsymbol{x})}{2\lambda_d}$ is the convolution between $\phi(\boldsymbol{x})$ and $\frac{p_d(\boldsymbol{x}) - p_g(\boldsymbol{x})}{2\lambda_d}$, which results:

$$D_p^*(\boldsymbol{x}) = \frac{1}{2\lambda_d}\int_{\mathcal{X}} \phi(\boldsymbol{x}-\boldsymbol{y})\left(p_d(\boldsymbol{y}) - p_g(\boldsymbol{y})\right)\, \mathrm{d}\boldsymbol{y}, \tag{11}$$

Including the family of homogeneous solutions $D_h^*(\boldsymbol{x}) = \langle \boldsymbol{a}, \boldsymbol{x}\rangle + \text{constant}$, the general solution becomes

$$D^*(\boldsymbol{x}) = D_p^*(\boldsymbol{x}) + \langle \boldsymbol{a}, \boldsymbol{x}\rangle + \text{constant}. \tag{12}$$

C.2 Optimal WGAN-FS Generator

The derivation of the optimal generator p_g^* proceeds along the lines of the first-variation analysis, taking into account the fact that the generator cost does not involve terms containing the derivatives of p_g. Consider the Lagrangian

$$\mathcal{L}_G = \int_{\mathcal{X}} (p_d(\boldsymbol{x}) - p_g(\boldsymbol{x}))D^*(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x},$$

where $D^*(\boldsymbol{x})$ is as given in Equation (12):

$$D^*(\boldsymbol{x}) = \frac{1}{\lambda_d}\int_{\mathcal{X}} \phi(\boldsymbol{x}-\boldsymbol{y})\left(p_d(\boldsymbol{y}) - p_g(\boldsymbol{y})\right)\, \mathrm{d}\boldsymbol{y} + \langle \boldsymbol{a}, \boldsymbol{x}\rangle + \text{ constant},$$

with $\phi(\boldsymbol{x}) = \kappa_n \|\boldsymbol{x}\|^{2-n}$, where $\boldsymbol{x} \in \mathbb{R}^n,\ n \geq 3$, $\kappa_n = \frac{1}{n(n-2)\mathfrak{v}(n)}$, $\mathfrak{v}(n)$ is the volume of the unit sphere in $\mathbb{R}^n$, and $\mathcal{X}$ is the convex hull of the supports of p_d and p_g. Denote the optimal generator as $p_g^*(\boldsymbol{x})$. Consider the perturbation $p_g^*(\boldsymbol{x}) + \epsilon\,\eta(\boldsymbol{x})$, where $\eta(x)$ is a family of compactly supported, absolutely integrable, infinitely differentiable functions that are identically zero at the boundaries of

$\mathcal{X}$. The first variation $\partial\mathcal{L}_G$ is given by

$$\begin{aligned}\partial\mathcal{L}_G &= \int_{\mathcal{X}}\int_{\mathcal{X}} \phi(\boldsymbol{y})\eta(\boldsymbol{x}-\boldsymbol{y})(p_g^*(\boldsymbol{x})-p_d(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\boldsymbol{x} \\ &\quad + \int_{\mathcal{X}} \Big(\langle \boldsymbol{a},\boldsymbol{x}\rangle - \big(\phi * (p_d - p_g^*)\big)(\boldsymbol{x})\Big)\eta(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \\ &= T_1 + T_2,\end{aligned}$$

where $\alpha_d = \dfrac{\lambda_d^*}{4}$. The term T_1 involves a convolution with a singular kernel $\phi(\boldsymbol{y})$, with the singularity at the origin. The integrals therefore have to be evaluated in the Cauchy principal-value sense. We make the interpretation explicit by defining:

$$\text{p.v.}\int_{\mathcal{X}}(\ \cdot\ )\,\mathrm{d}\boldsymbol{x} = \lim_{\xi\to 0}\int_{\mathcal{X}^\xi}(\ \cdot\ )\,\mathrm{d}\boldsymbol{x},$$

where $\mathcal{X}^\xi = \mathcal{X} - \mathcal{B}(0,\xi)$, which is formed by removing a ball of radius ξ centered at the origin. Recall that $\mathcal{X}$ is assumed to be compactly supported, and hence $\mathcal{X}^\xi$ is compactly supported as well. Consider η to be absolutely integrable over $\mathcal{X}^\xi$. Applying Fubini's theorem to T_1 yields

$$\begin{aligned}T_1 &= \lim_{\xi\to 0}\int_{\mathcal{X}^\xi}\phi(\boldsymbol{y})\int_{\mathcal{X}^\xi}\big(p_g^*(\boldsymbol{x})-p_d(\boldsymbol{x})\big)\,\eta(\boldsymbol{x}-\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}, \\ &= \lim_{\xi\to 0}\int_{\mathcal{X}^\xi}\int_{\mathcal{X}^\xi}\phi(\boldsymbol{y})\,\big(p_g^*(\boldsymbol{x}+\boldsymbol{y})-p_d(\boldsymbol{x}+\boldsymbol{y})\big)\,\eta(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}.\end{aligned}$$

Swapping the order of integration yields

$$T_1 = \lim_{\xi\to 0}\int_{\mathcal{X}^\xi}\eta(\boldsymbol{x})\,\big(\phi * \big(p_g^* - p_d\big)(\boldsymbol{x})\big)\ \mathrm{d}\boldsymbol{x},$$

since ϕ is radially symmetric. Substituting T_1 back into $\partial\mathcal{L}_G$, setting it to zero, and invoking the fundamental lemma of calculus of variations (cf. Appendix A), we obtain the condition

$$\big(\phi * (p_g^* - p_d)\big)(\boldsymbol{x}) = \frac{1}{2}\langle \boldsymbol{a},\boldsymbol{x}\rangle, \tag{13}$$

which the optimal generator p_g^* must satisfy. Applying the Laplacian operator Δ on both sides of Equation (13), we get

$$p_g^*(\boldsymbol{x}) = p_d(\boldsymbol{x}), \tag{14}$$

which is the desired optimality condition of the generator distribution, and is independent of the choice of the homogeneous component $D_h(\boldsymbol{x})$.

C.3 Optimal Lagrange Multiplier in WGAN-FS

Consider the Fourier-series (FS) discriminator $D_{FS}^*(\boldsymbol{x})$ in the multivariate case:

$$D_{FS}^*(\boldsymbol{x}) \approx \frac{1}{\lambda_{FS}^{*^2}}\left(\langle \boldsymbol{a},\boldsymbol{x}\rangle + \text{constant} + \sum_{\boldsymbol{m}\in\mathcal{M}}\big(\gamma_{\boldsymbol{m}}^r\cos(\omega_o\langle \boldsymbol{m},\boldsymbol{x}\rangle) + \gamma_{\boldsymbol{m}}^i\sin(\omega_o\langle \boldsymbol{m},\boldsymbol{x}\rangle)\big)\right).$$

Taking the derivative with respect to x_ℓ and squaring, we get:

$$\left(\frac{\partial D_{FS}^*}{\partial x_\ell}\right)^2 = \frac{1}{\lambda_{FS}^{*^2}}\left(a_\ell - \sum_{\boldsymbol{m}\in\mathcal{M}}\big(\gamma_{\boldsymbol{m}}^r\omega_o m_\ell\sin(\omega_o\langle \boldsymbol{m},\boldsymbol{x}\rangle) + \gamma_{\boldsymbol{m}}^i\omega_o m_\ell\cos(\omega_o\langle \boldsymbol{m},\boldsymbol{x}\rangle)\big)\right)^2.$$

Using the Cauchy-Schwartz inequality:

$$\left(\sum_{\ell=1}^n u_\ell.1\right)^2 \ \le\ n\sum_{\ell=1}^n u_\ell^2,$$

we obtain the following bound:

$$\left(\frac{\partial D^*_{FS}}{\partial x_\ell}\right)^2 \leq \frac{2|\mathcal{M}|+1}{\lambda^{*2}_{FS}}\left(a_\ell^2 + \sum_{\boldsymbol{m}\in\mathcal{M}} \omega_o^2 m_\ell^2 \left(\gamma_{\boldsymbol{m}}^{r^2}\sin^2(\omega_o\langle \boldsymbol{m},\boldsymbol{x}\rangle) + \gamma_{\boldsymbol{m}}^{i^2}\cos^2(\omega_o\langle \boldsymbol{m},\boldsymbol{x}\rangle)\right)\right),$$

where $|\mathcal{M}|$ is the cardinality of the set $\mathcal{M}$ of the selected harmonics. Summing over ℓ yields:

$$\|\nabla D^*\|_2^2 \leq \frac{2|\mathcal{M}|+1}{\lambda^{*2}_{FS}}\left(\|\boldsymbol{a}\|^2 + \sum_{\boldsymbol{m}\in\mathcal{M}} \omega_o^2\|\boldsymbol{m}\|^2 \left(\gamma_{\boldsymbol{m}}^{r^2}\sin^2(\omega_o\langle \boldsymbol{m},\boldsymbol{x}\rangle) + \gamma_{\boldsymbol{m}}^{i^2}\cos^2(\omega_o\langle \boldsymbol{m},\boldsymbol{x}\rangle)\right)\right)$$

$$\Rightarrow \|\nabla D^*\|_2^2 \leq \frac{2|\mathcal{M}|+1}{\lambda^{*2}_{FS}}\left(\|\boldsymbol{a}\|^2 + \sum_{\boldsymbol{m}\in\mathcal{M}} \left(\left(\tau_{\boldsymbol{m}}^i + \tau_{\boldsymbol{m}}^r\right) + \left(\tau_{\boldsymbol{m}}^i - \tau_{\boldsymbol{m}}^r\right)\cos(2\omega_o\langle \boldsymbol{m},\boldsymbol{x}\rangle)\right)\right),$$

$$\text{where } \tau_{\boldsymbol{m}}^r = \frac{1}{2}(\gamma_{\boldsymbol{m}}^r)^2\omega_o^2\|\boldsymbol{m}\|^2, \quad \text{and} \quad \tau_{\boldsymbol{m}}^i = \frac{1}{2}(\gamma_{\boldsymbol{m}}^i)^2\omega_o^2\|\boldsymbol{m}\|^2.$$

Enforcing the gradient-norm penalty: $\int_{\mathcal{X}} \left(\|\nabla D^*\|_2^2 - 1\right)\, \mathrm{d}\boldsymbol{x} = 0$, gives

$$0 \leq \int_{\mathcal{X}} \left((2|\mathcal{M}|+1)\left(\|\boldsymbol{a}\|^2 + \sum_{\boldsymbol{m}\in\mathcal{M}} \left(\left(\tau_{\boldsymbol{m}}^i + \tau_{\boldsymbol{m}}^r\right) + \left(\tau_{\boldsymbol{m}}^i - \tau_{\boldsymbol{m}}^r\right)\cos(2\omega_o\langle \boldsymbol{m},\boldsymbol{x}\rangle)\right)\right) - \lambda^{*2}_{FS}\right)\mathrm{d}\boldsymbol{x}.$$

Simplifying the above gives the condition on the optimal Lagrange multiplier:

$$\lambda^{*2}_{FS} \leq \frac{(2|\mathcal{M}|+1)}{|\mathcal{X}|}\int_{\mathcal{X}} \left(\|\boldsymbol{a}\|^2 + \sum_{\boldsymbol{m}\in\mathcal{M}} \left(\left(\tau_{\boldsymbol{m}}^i + \tau_{\boldsymbol{m}}^r\right) + \left(\tau_{\boldsymbol{m}}^i - \tau_{\boldsymbol{m}}^r\right)\cos(2\omega_o\langle \boldsymbol{m},\boldsymbol{x}\rangle)\right)\right)\, \mathrm{d}\boldsymbol{x}.$$

$$= (2|\mathcal{M}|+1)\left(\|\boldsymbol{a}\|^2 + \sum_{\boldsymbol{m}\in\mathcal{M}} \left(\tau_{\boldsymbol{m}}^i + \tau_{\boldsymbol{m}}^r\right) + \sum_{\boldsymbol{m}\in\mathcal{M}} \left(\left(\frac{\tau_{\boldsymbol{m}}^i - \tau_{\boldsymbol{m}}^r}{|\mathcal{X}|}\right)\int_{\mathcal{X}} \cos(2\omega_o\langle \boldsymbol{m},\boldsymbol{x}\rangle)\mathrm{d}\boldsymbol{x}\right)\right).$$

Given the data

$$\mathcal{D} = \{\boldsymbol{x}_k\} = \{\boldsymbol{x}_d, \text{ s.t. } \boldsymbol{x}_d \sim p_d\} \bigcup \{\boldsymbol{x}_g, \text{s.t. } \boldsymbol{x}_g \sim p_g\}$$

of cardinality $|\mathcal{D}| = N$, we can estimate the upper bound on λ^*_{FS} as follows:

$$\lambda^*_{FS} \leq \sqrt{(2|\mathcal{M}|+1)\left(\|\boldsymbol{a}\|^2 + \sum_{\boldsymbol{m}\in\mathcal{M}} \left(\tau_{\boldsymbol{m}}^i + \tau_{\boldsymbol{m}}^r\right) + \frac{1}{N}\sum_{k=1}^{N}\sum_{\boldsymbol{m}\in\mathcal{M}} \left(\tau_{\boldsymbol{m}}^i - \tau_{\boldsymbol{m}}^r\right)\cos(2\omega_o\langle \boldsymbol{m},\boldsymbol{x}_k\rangle)\right)}.$$

Recall that $\|\boldsymbol{a}\| = 1$ for the optimal discriminator to satisfy the gradient-norm penalty Ω_D when $p_g^* = p_d$ (cf. Section 3). In practice, the contribution of $\|\boldsymbol{a}\|$ was found to be negligible in comparison with the other terms. The worst-case choice for the Lagrange multiplier is

$$\lambda^*_{FS} = \sqrt{(2|\mathcal{M}|+1)\left(\sum_{\boldsymbol{m}\in\mathcal{M}} \left(\tau_{\boldsymbol{m}}^i + \tau_{\boldsymbol{m}}^r\right) + \frac{1}{N}\sum_{k=1}^{N}\sum_{\boldsymbol{m}\in\mathcal{M}} \left(\tau_{\boldsymbol{m}}^i - \tau_{\boldsymbol{m}}^r\right)\cos(2\omega_o\langle \boldsymbol{m},\boldsymbol{x}_k\rangle)\right)}.$$

D Additional Experimentation

In this appendix, we present additional experiments and results on univariate and multivariate synthetic Gaussian data, and on learning the image-space distributions with WGAN-FS. We also provide additional details on the evaluation metrics used.

D.1 Experimental Setup

We compare WGAN-FS with the following two categories of baselines: (i) WGAN and its variants with different penalties, such as the gradient penalty (WGAN-GP), Lipschitz penalty (WGAN-LP), Sobolev GAN and stable alternatives to GP, such as WGAN-R$_d$ and WGAN-R$_g$; and (ii) base WGAN with variations of the proposed gradient-norm penalty (GNP), evaluated empirically on sample

points drawn from the two datasets. WGAN-GNP implements the WGAN-GP algorithm with the GNP cost. While we compute the optimal Lagrange multiplier λ_d in closed-form in WGAN-FS, in Sobolev GANs, λ_d is optimized to maximize the discriminator loss through stochastic gradient-descent (Mroueh et al., 2018). Recently, multi-layer networks with periodic sinusoidal activations (SIREN) have been shown to achieve state-of-the-art performance in learning image, sound and wavefield representations (Sitzmann et al., 2020). We therefore adopt two variants of SIREN for the discriminator: (a) A three-layer fully connected network with $\sin$ activation, called WGAN-GNP (3S); and (b) A single-layer fully connected network with $\sin$ activation and the same number of nodes as terms in the Fourier-series expansion, called WGAN-GNP (1S). Training WGAN-GNP (1S) is equivalent to learning the Fourier coefficients in the WGAN-FS formulation.

Gaussian noise z that is input to the generator is sampled from the standard Gaussian $\mathcal{N}(0, 1)$. While WGAN-FS uses a closed-form Fourier-series discriminator, the baselines use a three-layer fully connected discriminator network with leaky ReLU activation. The batch size is 500. For the baseline techniques, each training step involves 5 iterations of the discriminator network optimization followed by one iteration of the generator. WGAN-FS, on the other hand, uses a single-shot discriminator during each training step. The Adam optimizer (Kingma & Ba, 2015) is used with a learning rate $\eta = 0.05$, and the exponential decay parameters for the first and second moments are $\beta_1 = 0.5$ and $\beta_2 = 0.999$, respectively. The implementation was carried out using TensorFlow 2.0 (Abadi et al., 2016).

D.2 Additional Experiments on 1-D and 2-D Gaussians

To begin with, we present results on learning 1-D and 2-D Gaussians and Gaussian mixtures with the WGAN-FS algorithm.

Accuracy of the Fourier-series approximation: For this experiment, Gaussian training data is drawn from $\mathcal{N}(10, 1)$. The fundamental period T is set to 7 in all the experiments. In Figure 2, we present the target distribution p_d and its Fourier-series approximation for various choices of truncation order M and batch size N to illustrate the trade-off between truncating the Fourier series at low frequencies, and the error in approximating high-frequency coefficients with sparse samples. We observe that, when M is small (e.g., $M = 5$), introducing additional samples does not improve the quality of the approximation. For larger M, (e.g., $M \geq 25$), we observe that, in line with the theory, the high-frequency terms have a larger variance in their estimate and require larger N to be estimated accurately. This is the statistical component of the error, which can be reduced by increasing N. The artifacts can be suppressed from the approximation by setting $N > M^{n+1}$ (for example, with $N = 500$ for $M = 10$ and $N = 1000$ for $M = 25$). We observe similar performance trade-offs in the case of learning a bimodal Gaussian mixture in 1-D, as shown in Figure 3. Additionally, when N and M are both small, the Fourier-series approximation fails to capture the smaller mode. Based on these observations, we expect WGAN-FS to perform relatively better with lower M even in the high-dimensional setting. Analytical bounds on the truncation and approximation error are derived in the Journal version (Asokan & Seelamantula, 2023).

Choosing the fundamental period T: We next present results on varying the assumed period T, given truncation order M and batch size N. Based on the previous experiments, we set $M = 10$ and $N = 100$. We consider the 1-D Gaussian learning scenario as above. The target is a Gaussian $\mathcal{N}(5, 1)$, while the noise distribution is $\mathcal{N}(0, 1)$. We compare results for various choices of the time period $T \in \{2, 5, 7, 11, 25, 75\}$. Figure 4 compares the quality of the Fourier-series approximation of the target distribution for each value of T. Since a Gaussian is infinitely supported, there will be aliasing in the Fourier representation no matter what the choice of the period is. In order to capture maximum area under the curve, to keep the aliasing error small, and to prevent the generator from latching on to an aliased version of the target density, we choose T to encompass 12σ supports of both the generator and the target densities in the fundamental period (for example, $T \geq 6$ for the standard normal distribution). A good choice of the fundamental period T is one that is centered around the generator distribution, but also encompasses the target distribution. For the scenario where the standard normal $\mathcal{N}(0, 1)$ is chosen as the noise distribution when learning a target $\mathcal{N}(\mu, \sigma)$ we observe that $T \approx \max\{6,\ \mu + 6\sigma\}$ results in a superior quality of the Fourier-series approximation of the target.

Figures 5(a) and (b) plot the Wasserstein-2 distance $\mathcal{W}^{2,2}$ and generator loss $\mathcal{L}_G$, respectively, as a function of iterations for various T. We observe that, for small T, the generator latches on to an

aliased version of the target, resulting in a large value for $\mathcal{W}^{2,2}$, although the loss $\mathcal{L}_G$ converges to zero. Choosing a large value of T makes the distribution appear like a spike (high-frequency) in the fundamental period and therefore, an accurate representation requires a larger value of M. For large M, although the Fourier-series approximation is not accurate, the generator samples converge to the desired target samples in terms of $\mathcal{W}^{2,2}$ and $\mathcal{L}_G$ by virtue of uniqueness of the Fourier representation for a given set of samples. Figure 5(c) shows the learnt discriminator for various choices of T. For small T, the learnt discriminator is unable to classify the target and generator distributions accurately. By virtue of the truncated Fourier-series approximation, the discriminator always learns a smooth approximation of the target classifier.

Convergence of the optimal Lagrange multiplier: We next illustrate the suitability of the optimal Lagrange multiplier λ^*_{FS} to serve as a proxy to measure convergence of the GAN generator during training. Figure 6 shows λ^*_{FS} and the Wasserstein-2 distance ($\mathcal{W}^{2,2}$) between p_d and p_g as a function of iterations. We observe that, for higher learning rates ($lr \approx 10^{-1}$), λ^*_{FS} does not converge to zero, which may be attributed to the fact that the $\mathcal{W}^{2,2}$ metric measures the convergence only between the first- and second-order statistics, while λ^*_{FS} measures the coefficient-wise convergence between the Fourier-series of p_d and p_g, which indirectly measures the L_2 error between the generator and target densities. This suggests that, while the models converge in the Wasserstein-2 sense for higher learning rates, convergence in the L_2 sense occurs for lower rates (here, $lr \leq 10^{-2}$). Based on these results, we set the learning rate to 10^{-3} for the generator in the subsequent experiments.

Experiments on 8-component Gaussian mixtures: In the 8-component GMM experiment, isotropic Gaussians are considered with standard deviation 0.05 and means lying in $[0, 1] \times [0, 1]$. The noise that is input to the generator is drawn from $\mathcal{N}(\mathbf{0}_{100}, \mathbb{I}_{100})$. The generator architecture for all WGAN models under consideration consists of three fully connected layers of 128, 64, and 32 nodes with LeakyReLU activation in each layer. The output layer has two nodes and a sigmoid activation. Figures 8(a) and (b) depict the $\mathcal{W}^{2,2}$ metric and KL divergence, respectively, as a function of iterations for the WGAN baseline models and the proposed WGAN-FS on the GMM learning task. The KL divergence is estimated parametrically by binning batches of samples to form histograms. The Wasserstein-2 distance is computed as a sample estimate using the publicly released *Python optimal transport* library (Flamary et al., 2021). We observe that, for the given choice of parameters, the baseline WGAN and WGAN-GP models latched on to different modes of the GMM at different stages of the optimization, failing to capture the entire distribution. We observe that WGAN-FS converges to lower values of the metrics compared with the baselines. Figure 7 shows the convergence of the generator distribution to the target data distribution in each case, while the associated heat-map represents the level-set of $D^*(\boldsymbol{x})$ at the given iteration. We observe that, during the initial iterations of training, WGAN-FS learns a significantly better representation of the underlying distributions compared with the baselines. This is evident from the fact that, while the baselines require optimizing a neural network for the discriminator, WGAN-FS provides the optimal discriminator for a given generator in closed form/single-shot at each iteration. Figure 8(c) compares the difference in performance of WGAN-FS with and without the homogeneous solution included. The generator optimization is independent of the homogeneous solution, with nearly identical performance in both cases, which is in accordance with the theoretical results.

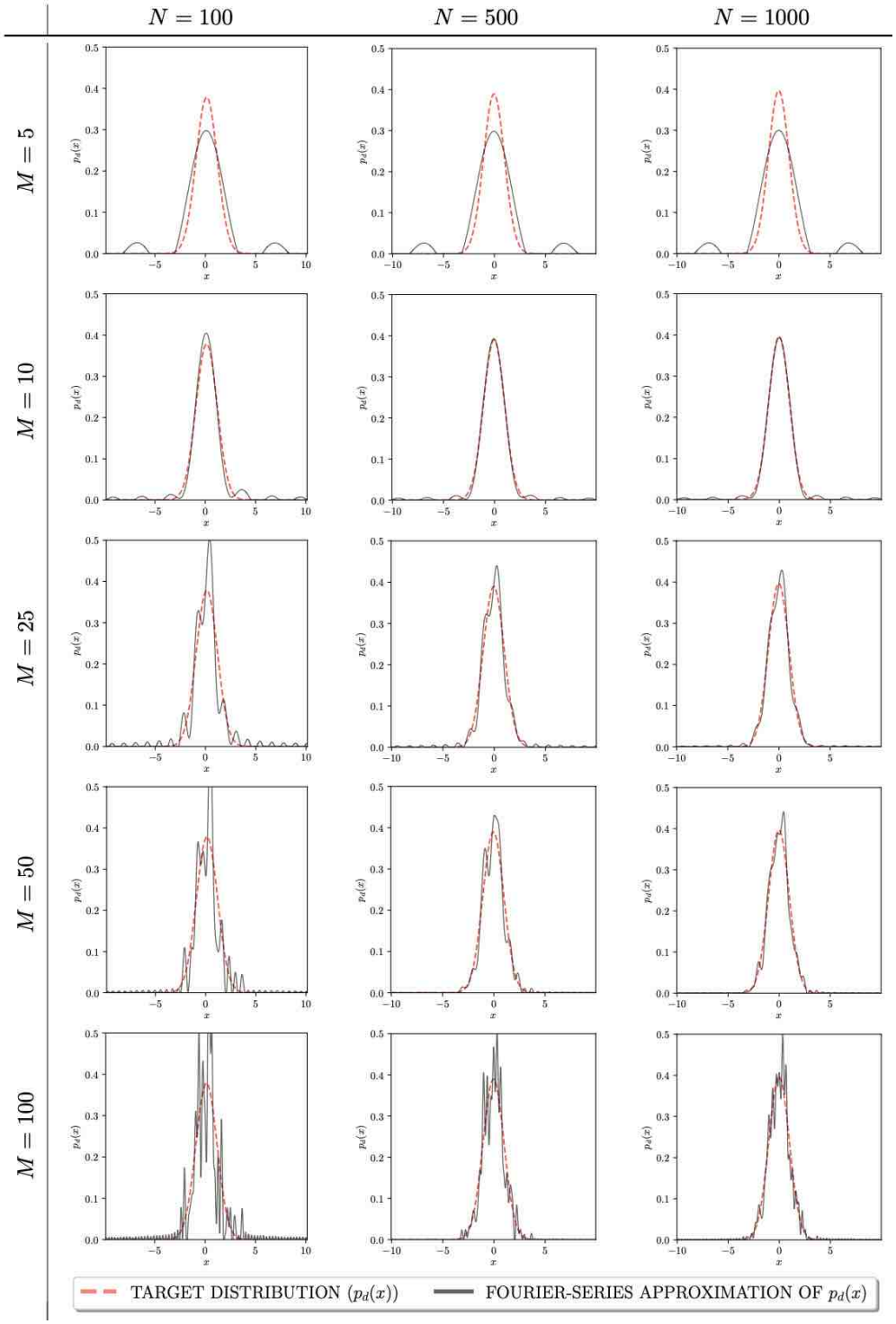

Figure 2: (Color online) Comparison of the quality of the Fourier-series approximation of a Gaussian $p_d(x)$ for various batch sizes N and truncation frequencies M.

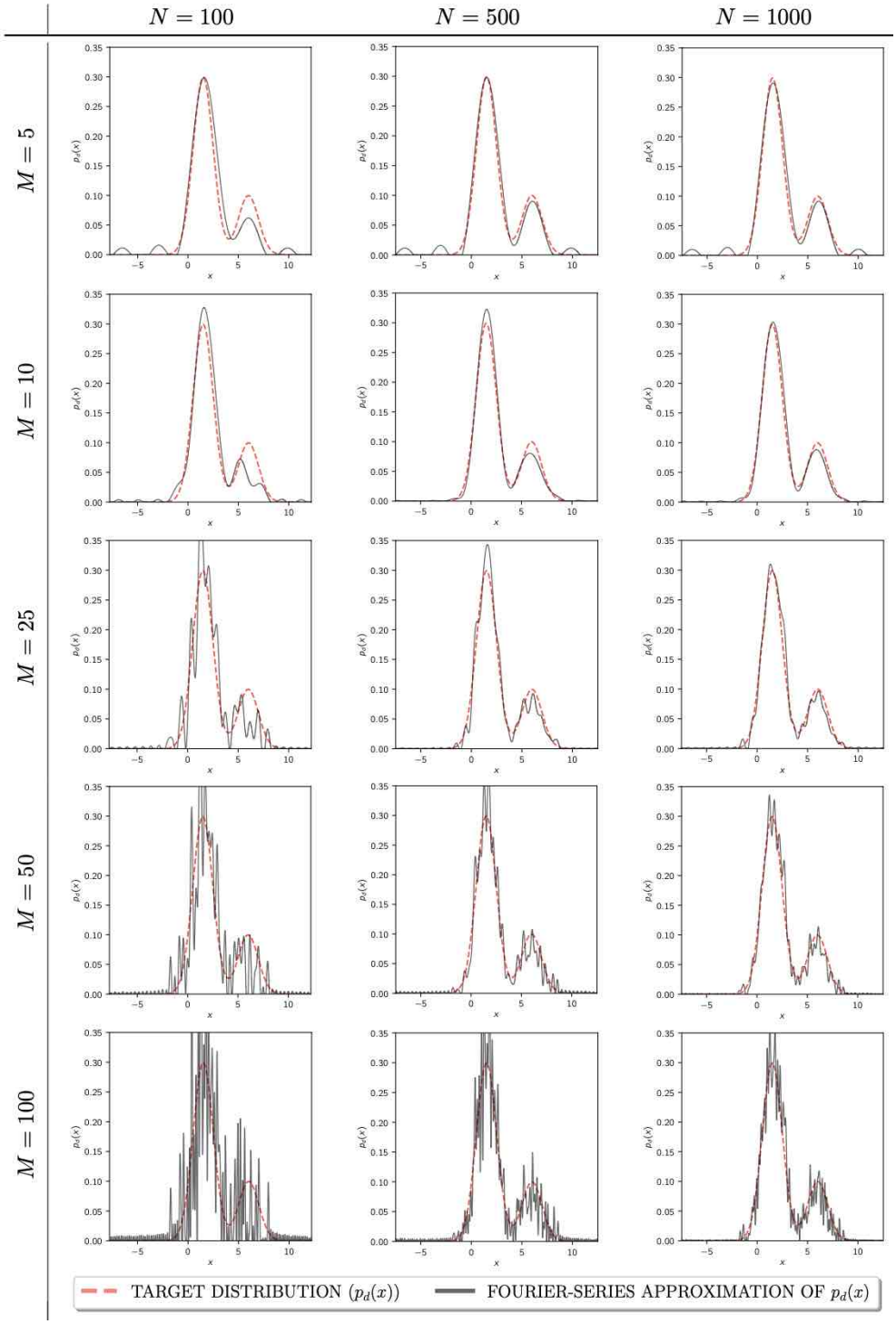

Figure 3: (Color online) Comparison of the quality of the Fourier-series approximation of a bimodal Gaussian p_d for various batch sizes N and truncation frequencies M.

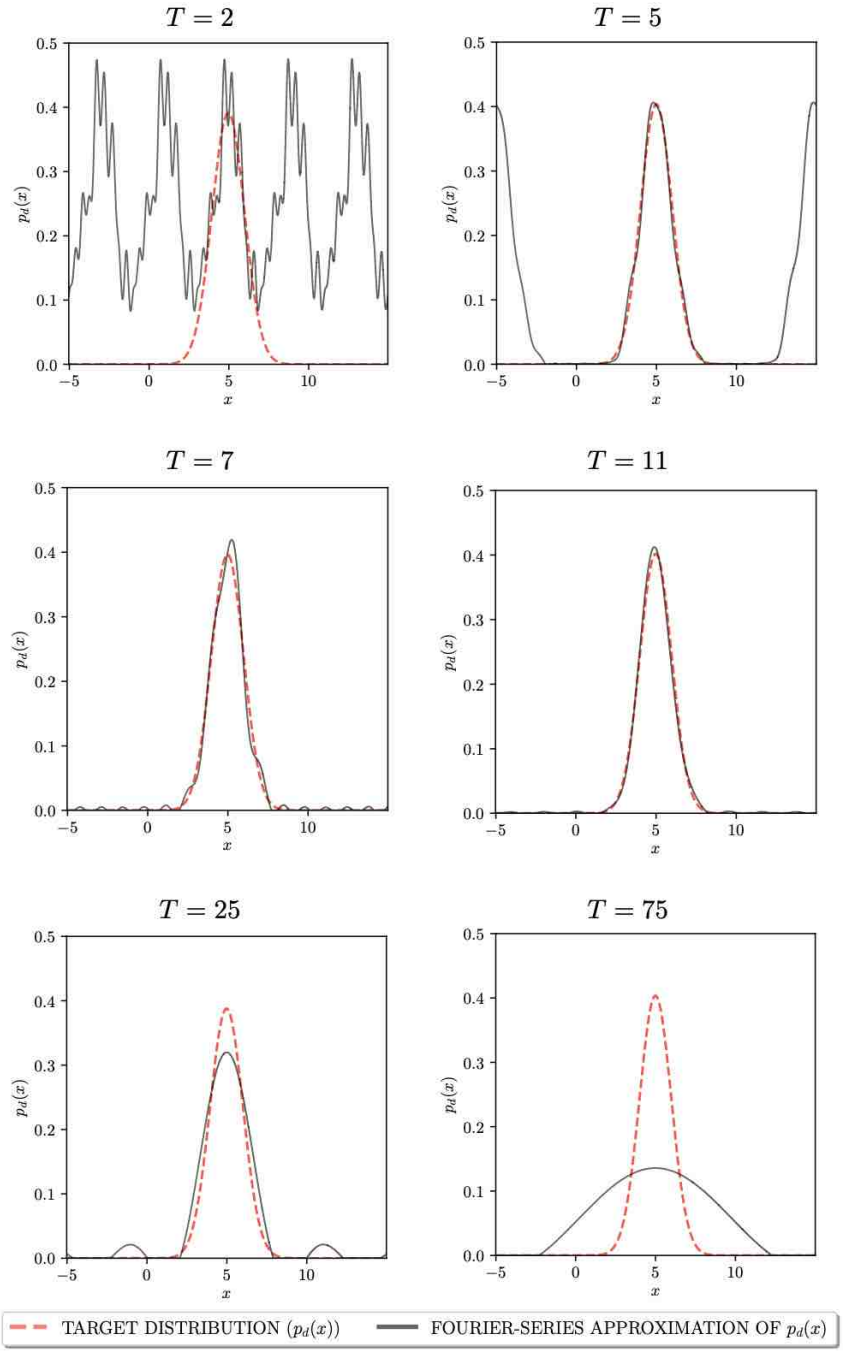

Figure 4: (Color online) Comparison of the quality of the 10-component Fourier-series approximation of a Gaussian $p_d(x)$ for various choices of the fundamental period T. Underestimating the time period results in aliasing, while overestimating it results in worse approximations of the distribution and requires additional high-frequency components in the expansion to improve upon the quality.

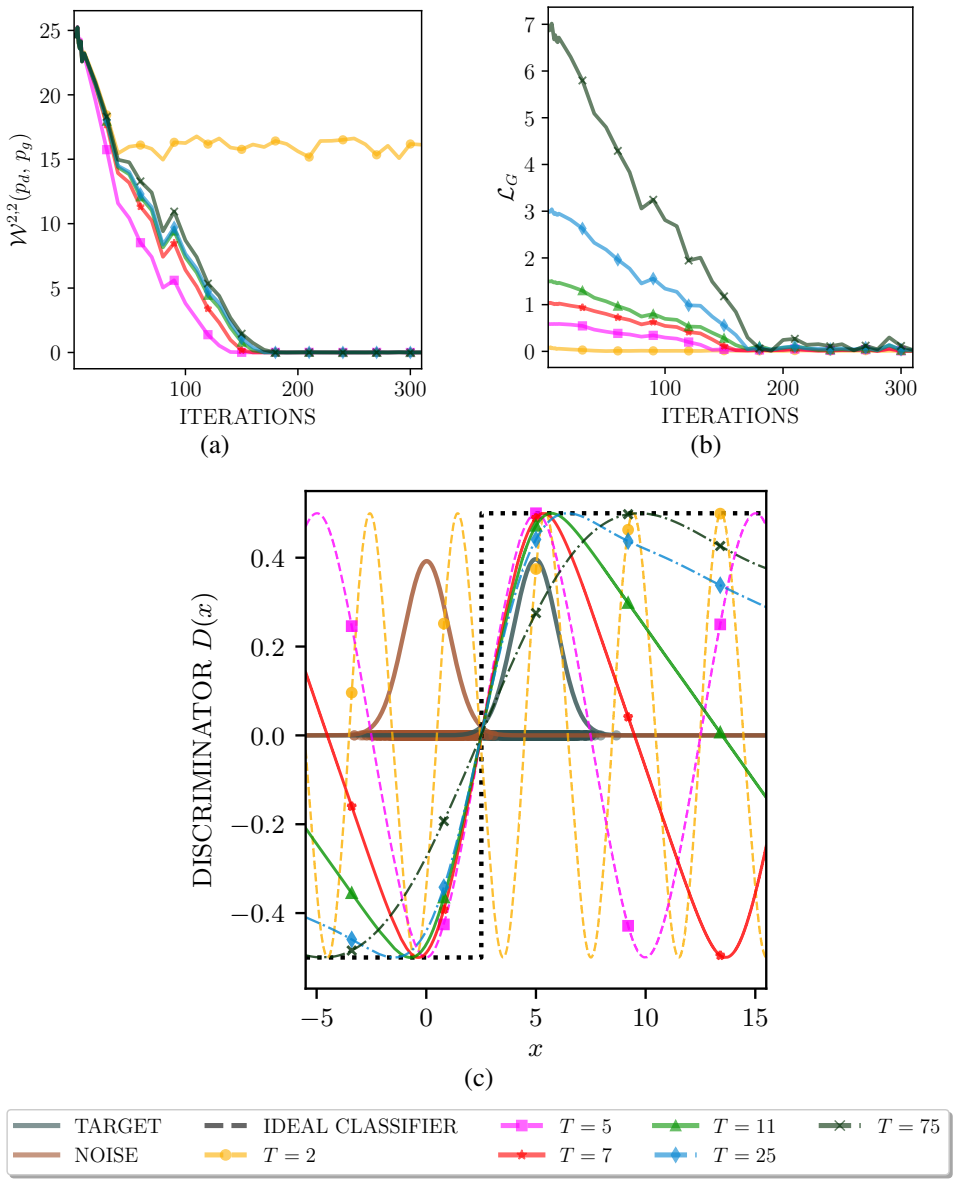

Figure 5: (Color online) Experiments on 1-D Gaussian data: Comparison of (a) Wasserstein-2 distance $\mathcal{W}^{2,2}(p_d, p_g)$; and (b) Generator loss $\mathcal{L}_G$ as a function of iterations when training WGAN-FS for various choices of T. For small T, the generator latches on to periodic replicas of the target, resulting in higher $\mathcal{W}^{2,2}$ values but low $\mathcal{L}_G$. (c) Comparison of the learnt discriminator when training WGAN-FS for various choices of T. WGAN-FS learns a smooth approximation of the true classifier for all T that contain 12σ windows of the generator and target distribution, thereby avoiding aliasing.

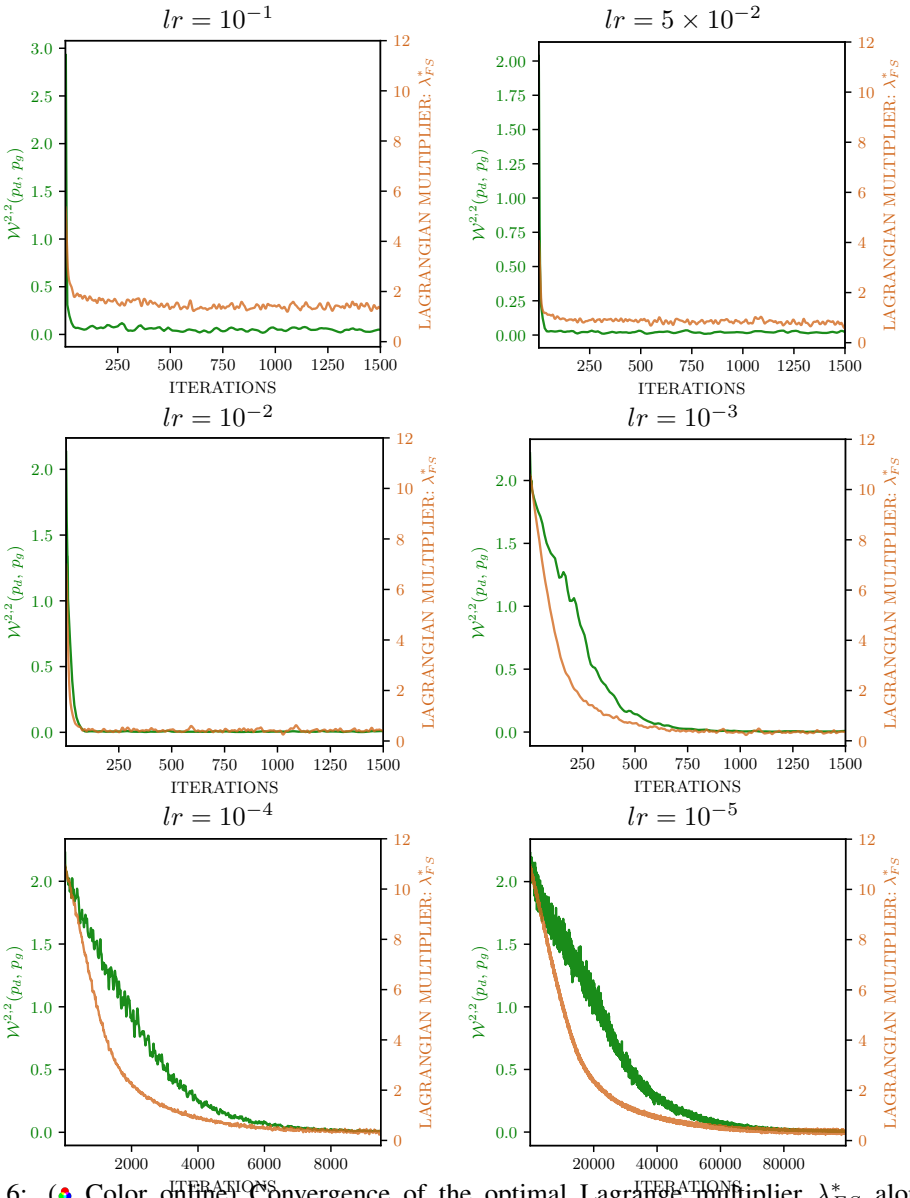

Figure 6: (Color online) Convergence of the optimal Lagrange multiplier λ^*_{FS} alongside Wasserstein-2 distance between p_d and p_g ($\mathcal{W}^{2,2}(p_d, p_g)$) for various learning rates. For higher learning rates, while the model appears to converge in the sense of $\mathcal{W}^{2,2}(p_d, p_g)$, which is a measure only up to second-order statistics, we observe from λ^*_{FS} that the distributions converge in the L_2 sense (the Fourier representation of p_g converging to that of p_d) only for learning rates lower than 10^{-2}. For very low rates (such as 10^{-5}), the convergence is not smooth. Therefore, we use learning rates in the range $[10^{-2}, 10^{-4}]$ in the subsequent experiments.

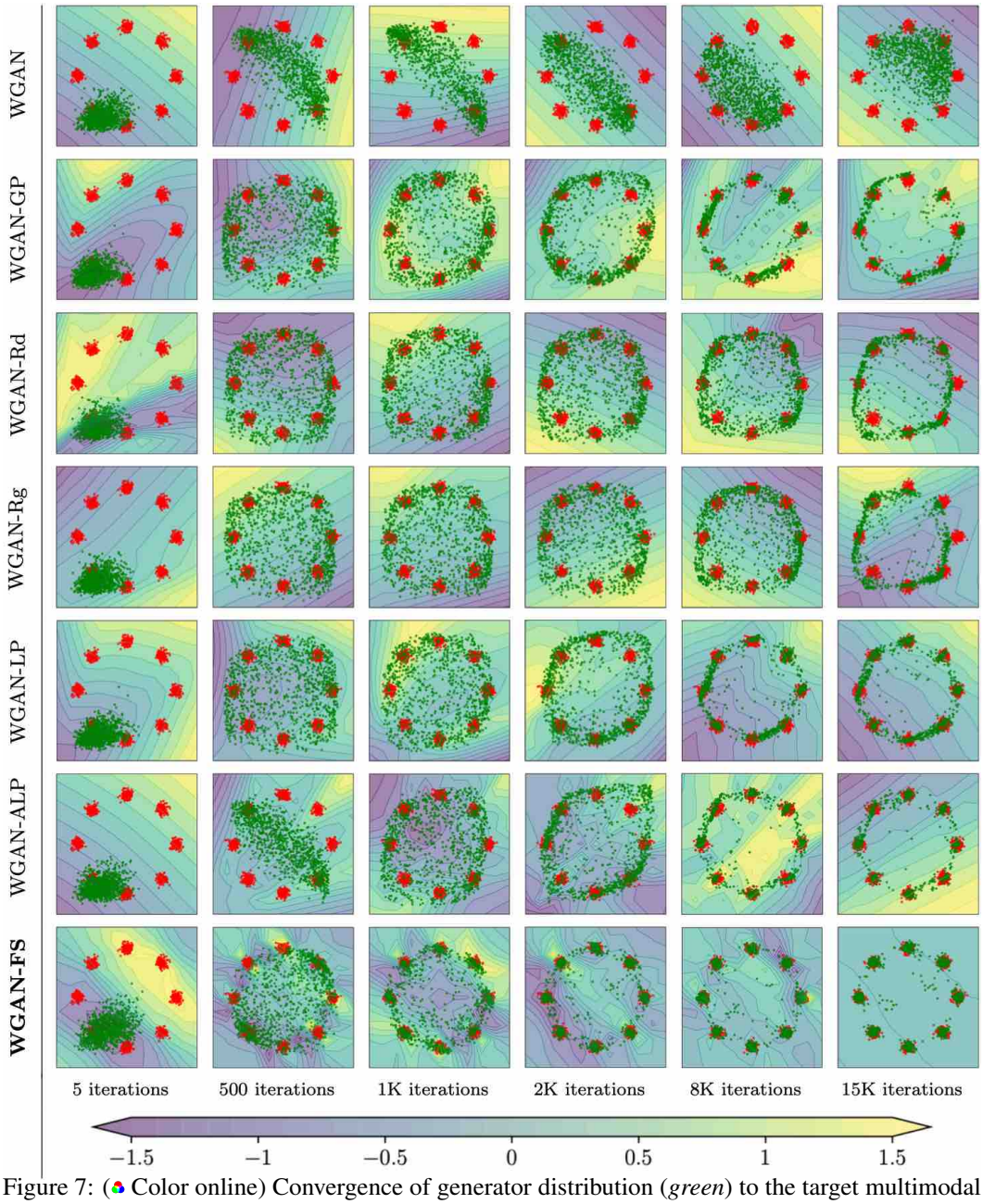

Figure 7: (● Color online) Convergence of generator distribution (*green*) to the target multimodal Gaussian data (*red*) on the considered WGAN variants. The heat map represents the values taken by discriminator. The ideal $D(\boldsymbol{x})$ is the one that takes larger values at locations where $p_d > p_g$ and vice versa, converging to a constant after p_g^* approaches p_d. The Fourier-series approximation of WGAN-FS approach leads to a better representation of the discriminator during the initial iterations than the baselines, leading to faster convergence. $1\text{K} = 1000$.

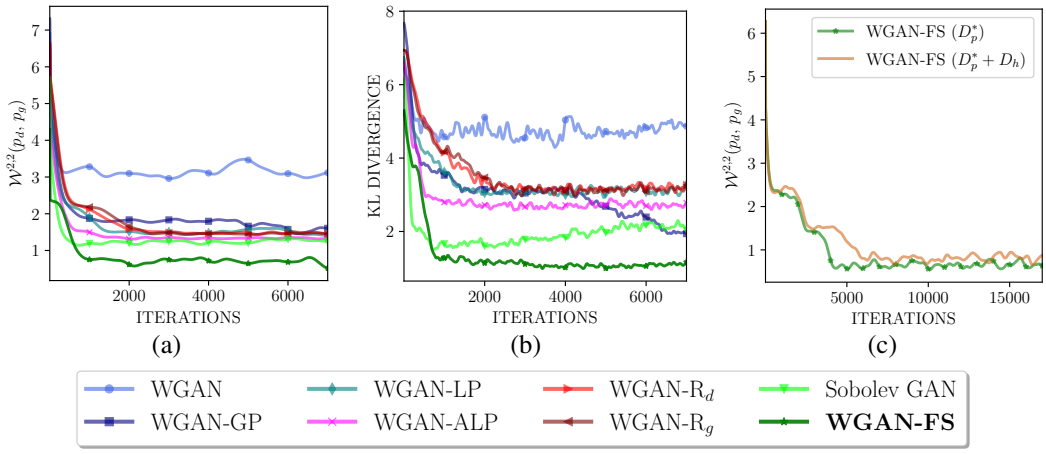

Figure 8: (Color online) Experiments on 2-D Gaussian-mixture data: Comparison of (a) Wasserstein-2 distance ($\mathcal{W}^{2,2}(p_d, p_g)$), and (b) Kullback-Leibler divergence between the data and generator distributions for WGAN-FS and baseline WGANs. WGAN-FS converges to a lower (better) value than the baselines in terms of both metrics. (c) Comparison of $\mathcal{W}^{2,2}(p_d, p_g)$ versus iterations for WGAN-FS with and without the homogeneous solution $D_h(\boldsymbol{x})$. The convergence of the WGAN-FS generator is relatively unaffected by the homogeneous component.

D.3 Experiments on n-dimensional Gaussians

We now present experimental results on learning multivariate Gaussian data with truncated Fourier-series expansions for WGAN-FS.

Experimental Setup: The experiments are conducted on n-D Gaussian data drawn from $\mathcal{N}(0.75\mathbf{1}_n, 0.2\mathbb{I}_n)$, where $\mathbf{1}_n$ denotes an n-dimensional vector with all entries equal to 1, and $\mathbb{I}_n$ is the n-dimensional identity matrix. The input to the generator is 100-D Gaussian noise. To simulate the scenario of training on real-world images with the WAE Encoder (Tolstikhin et al., 2018), the noise input is provided to a fully connected layer with $32 \times 32 \times 3$ nodes, whose output is reshaped to $(32, 32, 3)$. Subsequently, the reshaped noise vectors are provided as input to a network consisting of four convolution layers with 1024, 256, 128, and 64 filters in successive layers. The output of the convolution layers is flattened and provided to a fully connected layer with n output nodes. The learning rate is set to 10^{-2}, and batch size to $N = 100$. Recall that the Fourier-series expansion consists of two levels of approximation, one for the low-frequency part and the other for the high-frequency part. We consider all harmonics up to M_{low}, and a set of L distinct uniformly drawn/sampled harmonics between M_{low} and M_{high}. We pick $10 \leq n \leq 256$ to represent different latent space dimensions used in standard autoencoder architectures for images (Tolstikhin et al., 2018).

Results: Figure 9 shows the Wasserstein-2 metric $\mathcal{W}^{2,2}$, generator loss $\mathcal{L}_G$ and Lagrange multiplier λ_{FS}^* as a function of iterations, when training WGAN-FS to learn 10-D Gaussian data. We set $M_{low} = 2$ and $M_{high} = 10$. We experiment on multiple choices of the sample size: $L \in \{5, 10, 20, 100, 500, 1000, 10000, 25000\}$. We observe from Figure 9(a) that the model converges faster for smaller L (for example $L \leq 500$ in the experiments). However, as seen in Figure 9(b), for small L, the value of $\mathcal{L}_G$ is higher. From Figure 9(c), we see that for large L (such as $L > 10^3$), the convergence of the model in terms of λ_{FS}^* is slower. We attribute this to the slower convergence of the high-frequency components in the Fourier-series expansions due to increased variance in estimating these components for a given batch size N. This disparity is more pronounced when λ_{FS}^* is plotted on the logarithmic scale, as seen in Figure 9(b). We therefore chose $10^2 \leq L \leq 10^4$ to be a good compromise between achieving lower values of the generator loss and faster convergence of the model. The findings were similar when training the WGAN-FS model on 64-D and 128-D Gaussians (cf. Figures 10 and 11, respectively). We compare the performance of WGAN-FS for various n, given the sampling parameters $M_{low} = 2$, $M_{high} = 10$ and $L = 1000$. From Figure 12, we observe that, as n increases, both $\mathcal{W}^{2,2}$ and λ_{FS}^* exhibit poorer convergence (saturation to higher values). There is also increased jitter in the convergence of the loss and λ_{FS}^* as n increases.

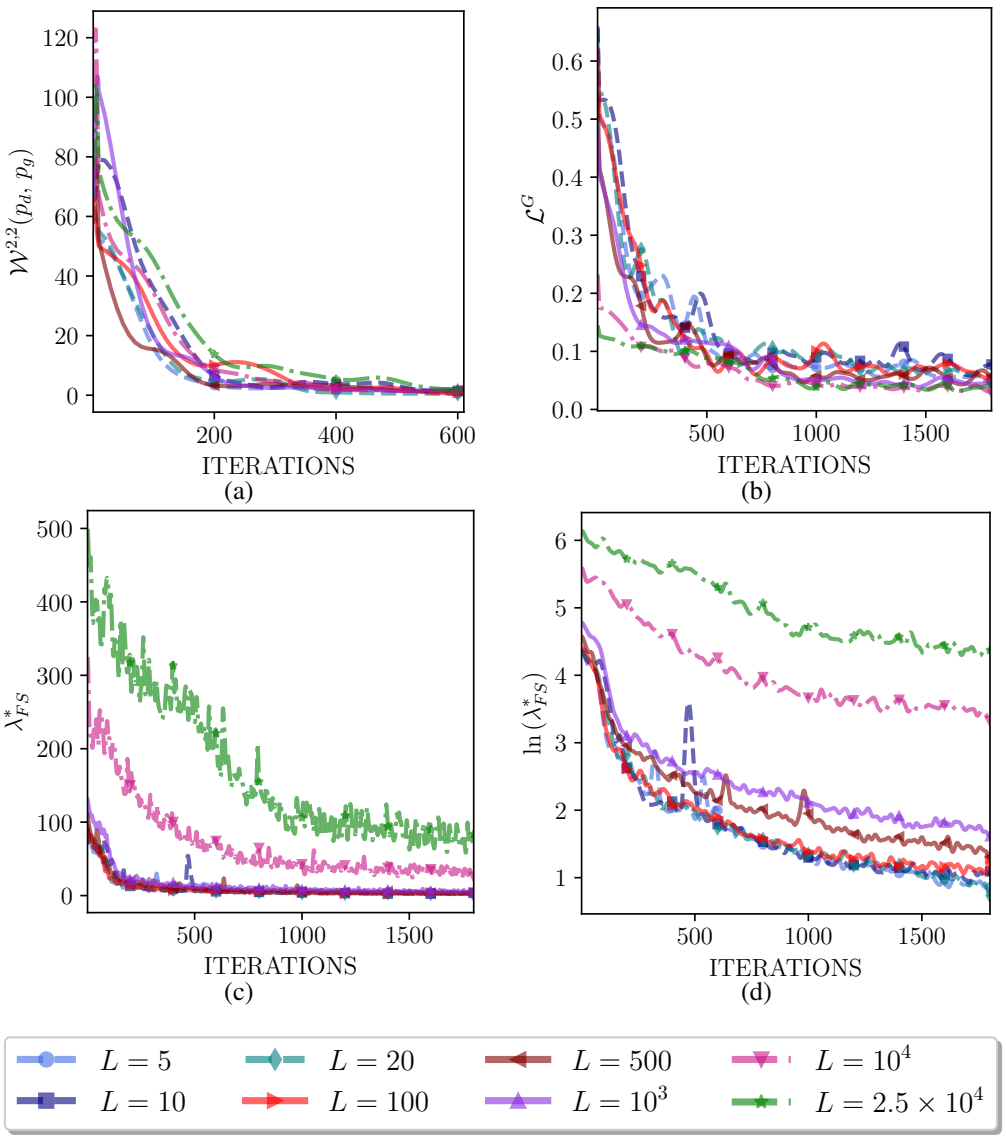

Figure 9: (Color online) Experiments on 10-D Gaussian data: Plots comparing the convergence of: (a) Wasserstein-2 distance $\mathcal{W}^{2,2}$; (b) Generator loss $\mathcal{L}_G$; (c) Optimal Lagrange multiplier λ^*_{FS}, and (d) the natural logarithm of λ^*_{FS} as a function of iterations when training WGAN-FS with L randomly sampled high-frequency components. The convergence is slower for large L as the error in estimating the coefficients increases with an increase in the number of high frequency terms.

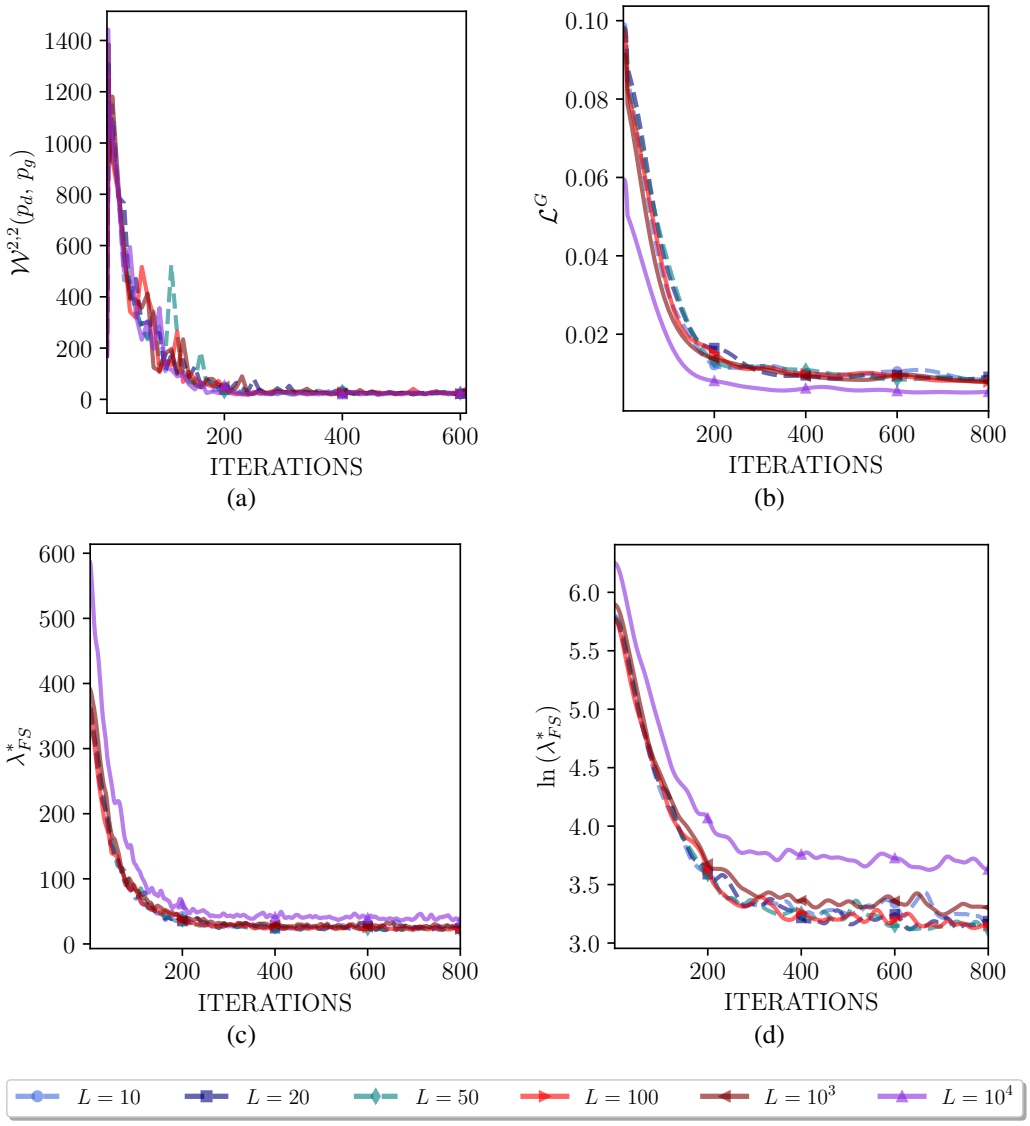

Figure 10: (Color online) Experiments on 64-D Gaussian data: Plots comparing the convergence of: (a) Wasserstein-2 distance $\mathcal{W}^{2,2}$; (b) Generator loss $\mathcal{L}_G$; (c) Optimal Lagrange multiplier λ^*_{FS}, and (d) the natural logarithm of λ^*_{FS} when training WGAN-FS on 64-dimensional Gaussian data for various number of sampled high-frequency coefficients, L. We observe that λ^*_{FS} converges to a worse (higher) value for larger L, while Wasserstein-2 distance $\mathcal{W}^{2,2}(p_d, p_g)$ and generator loss $\mathcal{L}_G$ are worse for small L. Setting L to be around 10^3 is a viable compromise.

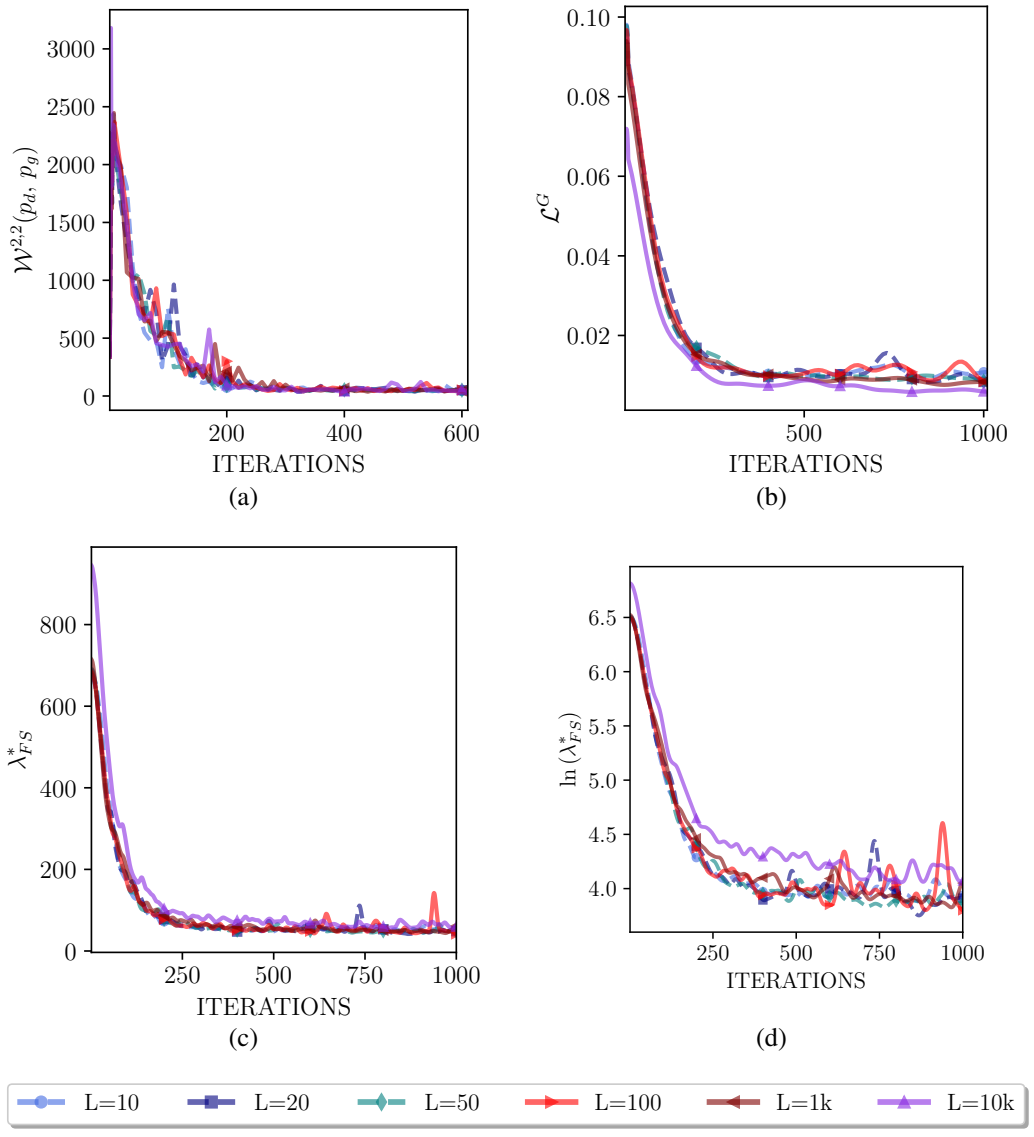

Figure 11: (Color online) Experiments on 128-D Gaussian data: Plots comparing the convergence of: (a) Wasserstein-2 distance $\mathcal{W}^{2,2}$; (b) Generator loss $\mathcal{L}_G$; (c) Optimal Lagrange multiplier λ^*_{FS}, and (d) the natural logarithm of λ^*_{FS} when training WGAN-FS on 128-dimensional Gaussian data for various number of sampled high-frequency coefficients, L. We observe that the models converge to worse (higher) values of λ^*_{FS} as L increases. This suggests that Fourier-series-based discriminator performs better when fewer high-frequency components are included in the approximation.

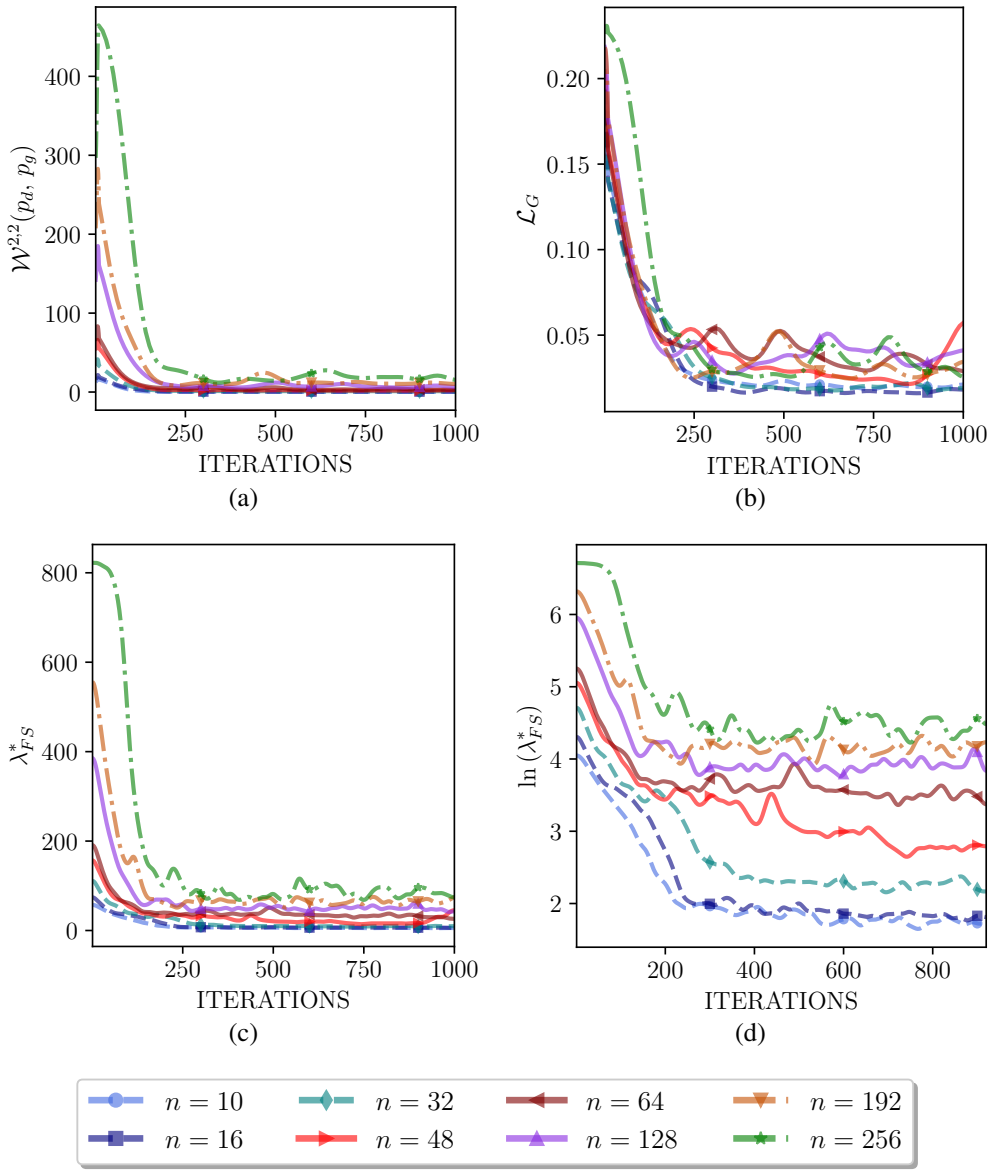

Figure 12: (Color online) Plots comparing the convergence of (a) Wasserstein-2 distance $\mathcal{W}^{2,2}(p_d, p_g)$, (b) Generator loss $\mathcal{L}_G$, (c) the optimal Lagrange multiplier λ^*_{FS}, and (d) the natural logarithm of λ^*_{FS} when training WGAN-FS on n-dimensional data, for various n. Across all three metrics, we observe that the models converge to worse (higher) values as the dimensionality of the data increases. This suggests that Fourier-series-based discriminator performs better on lower-dimensional latent-space matching.