Bridging the Gap Between Coulomb GAN and Gradient-regularized WGAN

Anonymous Author(s)
Affiliation
Address
email

Abstract

Generative adversarial networks (GANs) are essentially a min-max game between the discriminator and a generator. Coulomb GANs have a closely related formulation where the generator minimizes the potential difference between real (negative) and fake (positive) *charge densities*, wherein the discriminator approximates a low-dimensional Plummer kernel centered around the samples. Motivated by links between electrostatic potential theory and the Poisson partial differential equation (PDE), we consider the underlying functional optimization in Coulomb GAN and show that the associated discriminator is the optimum of a first-order gradient-regularized Wasserstein GAN (WGAN) cost. Subsequently, we show that, within the regularized WGAN setting, the optimal discriminator is the Green's function to the Poisson PDE, which corresponds to the Coulomb potential. As an alternative to training a discriminator in either WGAN or Coulomb GAN, we demonstrate, by means of synthetic data experiments, that the closed-form implementation of the optimal discriminator leads to a superior performance of the GAN generator.

1 Introduction

Generative adversarial networks (GANs) (Goodfellow et al., 2014) have gained immense popularity over the past decade due to their ability to generate realistic images, requiring low sampling times (Karras et al., 2019, 2020, 2021). GANs are a min-max game, where a generator network G is trained to learn the underlying distribution p_d of a target dataset with the aid of a discriminator network D. The generator is essentially a transformation of a noise distribution p_z, typically Gaussian, into the push-forward distribution $p_g = G_{\#}(p_z)$.

Wasserstein GANs (WGAN), introduced by Arjovsky et al. (2017) consider a discriminator that approximates a distance (called integral probability metric or IPM) between the target and generator distributions, which in turn, induces a constraint class from which the discriminator function is drawn. For example, WGANs consider Lipschitz-1 constraints, while WGAN-GP (Gulrajani et al., 2017) and Sobolev GANs (Mroueh et al., 2018) consider functions with bounds on the norm of the gradient. Subsequent variants proposed by Gulrajani et al. (2017); Bellemare et al. (2017); Mescheder et al. (2018) and Adler & Lunz (2018) also consider gradient-based regularizers to improve training stability. However, the min-max nature of the optimization makes it difficult to derive convergence guarantees on the GAN training, either in the parametric sense (Li et al., 2017b), or in the functional sense (Unterthiner et al., 2018).

Although the WGAN loss is a popular choice in most applications (Kang et al., 2022), in an effort to derive GANs with guaranteed convergence to the Nash equilibrium of the game, Unterthiner et al. (2018) consider a discriminator network that approximates the Plummer kernel evaluated at the real and generated sample points. They showed that, given the sample approximation of the

Submitted to the DLDE Workshop in the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Plummer-kernel discriminator, the GAN generator converges to the desired optimum, *i.e.* $p_g^* = p_d$. However, the links between the Coulomb GAN and IPM based losses was not explored and the underlying optimization problem that the discriminator solves is unclear. We therefore wish to ask the following question in this paper – How do the Plummer/Coulomb family of kernels link to classical IPM GAN optimization?

1.1 The Proposed Approach

We analyze the optimization problem in Coulomb GANs considering the Plummer kernel, Akin to potential theory, the optimal Coulomb GAN discriminator arises naturally as the consequence of gradient-regularization, more precisely, as a solution to the gradient-regularized Wasserstein GAN (WGAN). We analyze the gradient-regularized WGAN cost in a variational setting, and show that the optimal WGAN discriminator solves a Poisson PDE, whose Green's function gives rise to the Coulomb GAN's Plummer-kernel based potential function. Unlike in Coulomb GANs, where a discriminator network is used to approximate the kernel function, we implement the optimum as is, and demonstrate on synthetic data, that the closed form discriminator performs superior to both gradient-penalized WGANs, and the kernel-approximating Coulomb GANs. As a future scope of this work, one could solve the Poisson PDEs in high-dimensional spaces through appropriate numerical solvers, similar to the approaches used in Diffusion models (Ho et al., 2020) and score-matching approaches (Song & Ermon, 2019).

2 The Coulomb GAN Discriminator

Unterthiner et al. (2018) consider the task of generative learning, where the objective is to minimize the difference between the data and generator distributions, $\rho(\boldsymbol{x}) = p_g(\boldsymbol{x}) - p_d(\boldsymbol{x})$, given data $\boldsymbol{x} \in \mathbb{R}^n$. Motivated by the connection to electrostatics, they consider the potential function associated with the difference, and a chosen kernel κ, given by:

$$\Phi(\boldsymbol{x}) = \int_{\mathcal{X}} \rho(\boldsymbol{y})\kappa(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\boldsymbol{y}, \quad \text{where} \quad \kappa(\boldsymbol{x},\boldsymbol{y}) = \left(\sqrt{\|\boldsymbol{x}-\boldsymbol{y}\|^2+\epsilon^2}\right)^{-d}, \tag{1}$$

where $d \leq n-2$, $\mathcal{X}$ denotes the support of the discriminator, typically the convex hull encompassing the supports of p_d and p_g, and ϵ is a small non-zero constant used to eliminate the singularity in κ. For $d = n-2$, κ represents the Plummer kernel. Intuitively, the potential in Eq. (1) induces a field over $\mathcal{X}$ where the generator is forced to generate samples that are repelled from those samples deemed fake by the discriminator, while being attracted towards those classified as reals. In practice, Φ is approximated by a sample average, computed over a mini-batch, and is given by:

$$\hat{\Phi}(\boldsymbol{x}) = \frac{1}{N_g}\sum_{j=1}^{N_g}\kappa(\boldsymbol{x},\boldsymbol{g}_j) - \frac{1}{N_d}\sum_{i=1}^{N_d}\kappa(\boldsymbol{x},\boldsymbol{d}_i), \tag{2}$$

where $\boldsymbol{d}_i$ and $\boldsymbol{g}_j$ represent the N_d and N_g samples drawn from p_d and p_g, respectively. The Coulomb GAN discriminator is trained to approximate Φ in the least-squares sense, while the generator is trained to maximize D, the approximation of $\hat{\Phi}$, for its generated samples, *i.e.,*

$$\mathcal{L}_D^C(D;G) = \frac{1}{2}\,\mathbb{E}_{\boldsymbol{x}\sim p_x}\left[\left(D(\boldsymbol{x}) - \hat{\Phi}(\boldsymbol{x})\right)^2\right], \quad \text{and} \quad \mathcal{L}_G^C(G;D) = -\frac{1}{2}\,\mathbb{E}_{\boldsymbol{z}\sim p_z}\left[D(G(\boldsymbol{z}))\right],$$

where $p_x(\boldsymbol{x}) = \frac{1}{2}\int \mathcal{N}(\boldsymbol{x};G(\boldsymbol{z}),\varepsilon\mathbb{I})\,p_z(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} + \frac{1}{2}\int \mathcal{N}(\boldsymbol{x};\boldsymbol{y},\varepsilon\mathbb{I})\,p_d(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}$ denotes a distribution whose samples are drawn from a mix of p_d and p_g, and corrupted with Gaussian noise of a sufficiently small variance $\varepsilon\mathbb{I}$. Unlike in the standard GAN and WGAN formulations, Unterthiner et al. (2018) showed that the alternating minimization of $\mathcal{L}_D^C$ and $\mathcal{L}_G^C$ results in attaining the unique global optimum (D^*, G^*), where $p_g^* = G_\#^*(p_z) = p_d$, the desired data distribution. However, the nonlinearities introduced by the choice of κ, and the slow learning of the discriminator caused by the least-squares cost, mandate the use of small learning rates for training the two networks. Furthermore, the link between the Coulomb GAN discriminator cost $\mathcal{L}_D^C$ and other IPM based GAN optimization is not been fully explored. In this paper, we show that Coulomb GANs can be linked to the WGANs through the solution to the Poisson PDE.

3 Gradient-regularized WGAN and the Plummer Kernel

Arjovsky et al. (2017) presented the GAN learning problem as one of *optimal transport*, where the critic (or discriminator) minimizes the earth-mover's distance or Wasserstein-1 distance between p_d and p_g. Through the Kantorovich-Rubinstein duality, they defined the WGAN discriminator loss as $\mathcal{L}_D^W = \mathbb{E}_{\boldsymbol{x}\sim p_g}[D(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim p_d}[D(\boldsymbol{x})]$ and the generator loss as $\mathcal{L}_G^W = -\mathcal{L}_D^W$, such that $D(\boldsymbol{x})$ is Lipschitz-1. Gulrajani et al. (2017) approximate the Lipschitz-1 constraint though a gradient penalty $\mathbb{E}_{\boldsymbol{x}\sim p_{\text{int}}}\left[(\|\nabla D\left(\boldsymbol{x}\right)\|_2 - 1)^2\right]$, where p_{int} is an interpolated distribution between p_d and p_g. Subsequent works (Kodali et al., 2017; Petzka et al., 2018; Terjék, 2020) have replaced p_{int} with alternatives that resulted in stabler training. For example, Mescheder et al. (2018) and Mroueh et al. (2018) consider $\Omega_D^R = \mathbb{E}_{\boldsymbol{x}\sim p_r}\left[\|\nabla D(\boldsymbol{x})\|_2^2\right]$ for various choices of p_r.

WGANs, PDEs and the Coulomb Kernel: Mroueh et al. (2018) showed that the optimal solutions in Sobolev GANs solve a family of Fokker-Plank PDEs. However, they do not invert the general Fokker-Plank operator due to computational constraints. Recall that the Plummer-like kernel, with $\epsilon = 0$ represent the radially-symmetric Coulomb kernel, given by

$$\psi_n(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} |x - y| & \text{for } n = 1 \\ -\ln\left(\|\boldsymbol{x} - \boldsymbol{y}\|\right) & \text{for } n = 2 \\ \|\boldsymbol{x} - \boldsymbol{y}\|^{2-n} & \text{for } n = 3, 4, 5, \cdots \end{cases} \tag{3}$$

While generative moment matching networks (GMMNs) (Li et al., 2015) and maximum mean discrepancy GANs (MMD-GANs) (Li et al., 2017a) consider Gaussian and inverse multi-quadric kernels in the context of GAN learning, the Coulomb kernels have been extensively studied in interpolation literature (Aronszajn et al., 1983; Iske, 2004; Fasshauer, 2007). They are known to arise in mesh-free gradient-regularized least-squares interpolation, where the smoothness constraint is enforced over $\mathcal{X}$. Motivated by these connections, we consider the IPM GAN discriminator cost with a gradient-norm constraint, given by

$$\mathcal{L}_D = \mathbb{E}_{\boldsymbol{x}\sim p_g}[D(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim p_d}[D(\boldsymbol{x})] + \lambda_d \int_{\mathcal{X}} \|\nabla D(\boldsymbol{x})\|_2^2 \, \mathrm{d}\boldsymbol{x} \ . \tag{4}$$

The above penalty can be seen as a special case of Ω_D^R with p_r set to the uniform measure over $\mathcal{X}$. The following theorem links the discriminator in gradient-regularized GANs to the Coulomb GAN.

Theorem 3.1. *The **optimal discriminator** that minimizes the $\mathcal{L}_D$ solves the following Poisson PDE:*

$$\Delta D(\boldsymbol{x}) = \frac{\rho(\boldsymbol{x})}{2\lambda_d}, \ \forall\, \boldsymbol{x} \in \mathcal{X}, \tag{5}$$

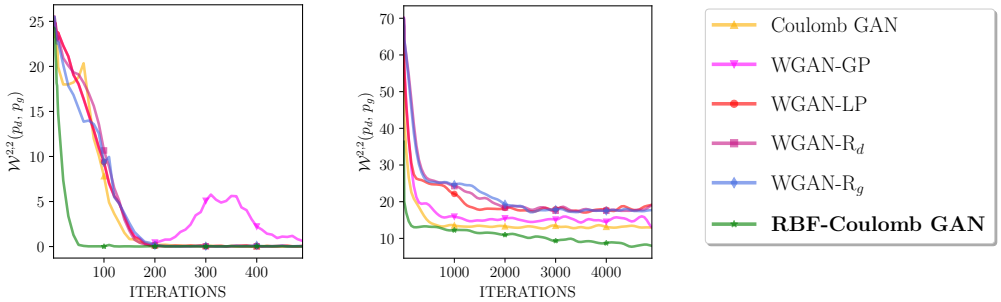
where Δ is the Laplacian operator. The particular solution, $D_p^*(\boldsymbol{x})$, is given by

$$D_p^*(\boldsymbol{x}) = \frac{1}{2\lambda_d}\left(\rho * \psi_n\right)(\boldsymbol{x}), \tag{6}$$

which is a multidimensional convolution with the Coulomb kernel, which in turn is the fundamental solution to the Laplace equation: $\Delta\psi_n(\boldsymbol{x}) = \mathrm{C}_n\, \delta(\boldsymbol{x})$, for some constant C_n.

Proof. The discriminator PDE can be derived through the *Calculus of Variations* applied to the cost in Eq. (4). PDEs of this type have been extensively researched and Evans (2010)'s book on *Partial Differential Equations* is a good reference on this topic. The details are provided in Appendix A. □

Eq. (5) indeed mimics results from electrostatics, and as observed in Coulomb GANs, the optimal discriminator is the potential function, with the generator and target distributions representing positive and negative charge densities, respectively. Additionally, these results show that optimizing the discriminator in GANs is closely linked to solving PDEs. While Theorem 3.1 provides the particular solution to the discriminator, the general solution also includes the homogeneous component, *i.e.,* solutions to $\Delta f(\boldsymbol{x}) = 0$, that belong to the space $\mathcal{P}_1^n$ of first-degree polynomials $P(\boldsymbol{x}) = \langle \boldsymbol{a}, \boldsymbol{x}\rangle + a_0$. The general solution to the discriminator PDE is $D^*(\boldsymbol{x}) = D_p^*(\boldsymbol{x}) + P(\boldsymbol{x})$. The exact choice of the polynomial $P(\boldsymbol{x}) \in \mathcal{P}_1^n$ depends on the boundary conditions and will be discussed shortly. The multidimensional convolution in a high-dimensional space makes Eq. (6) impractical. The following lemma presents an implementable form of the optimal discriminator.



(a) (b)

Figure 1: The Wasserstein-2 distance versus iterations on learning (a) a 2-D Gaussian; and (b) a 2-D Gaussian mixture; for various WGAN, Coulomb GAN and RBF-Coulomb GAN variants. The RBF-Coulomb GAN discriminator leads to a superior generator, compared to the baselines.

Lemma 3.2. *The optimal discriminator $D_p^*(\boldsymbol{x})$ given in Eq.* (6) *can be approximated through the following sample estimate:*

$$\tilde{D}_p^*(\boldsymbol{x}) = \frac{1}{2N_g\lambda_d} \sum_{\boldsymbol{g}_j \sim p_g} \psi_n(\boldsymbol{x} - \boldsymbol{g}_j) - \frac{1}{2N_d\lambda_d} \sum_{\boldsymbol{d}_i \sim p_d} \psi_n(\boldsymbol{x} - \boldsymbol{d}_j), \tag{7}$$

where ψ_n is as described in Eq. 3.

Lemma 3.2 shows that the sample approximation of the gradient-regularized WGAN discriminator equals the potential function described in Eq. 1 for $\epsilon = 0$. As opposed to training the GAN discriminator on $\mathcal{L}_D$ or $\mathcal{L}_D^C$ with gradient-descent, we directly implement $\hat{D}^*(\boldsymbol{x}) = \hat{D}_p^*(\boldsymbol{x}) + \langle \boldsymbol{a}, \boldsymbol{x} \rangle + a_0$. $\hat{D}_p^*(\boldsymbol{x})$ can be constructed using a radial basis function (RBF) network with pre-determined weights. Preliminary experiments indicate that the choice of the polynomial does not affect the generator optimization. Choosing $\boldsymbol{a} = \boldsymbol{0}$ and $a_0 = 0$ results in the Coulomb GAN solution.

4 Experimental Validation

To validate using a RBF discriminator, we present preliminary experiments on learning 2-D unimodal and multimodal Gaussian data. We compare the proposed approach, called *RBF-Coulomb GAN*, against the baseline Coulomb GAN, and gradient-regularized WGAN variants such as WGAN-GP (Gulrajani et al., 2017) , WGAN-R$_d$ and WGAN-R$_g$ (Mescheder et al., 2018), and WGAN-LP (Petzka et al., 2018). Network parameters and data specifications are described in Appendix C. Performance comparisons, using the Wasserstein-2 distance $\left(\mathcal{W}^{2,2}(p_d, p_g)\right)$, are presented in Figures 1(a) and (b). In both the scenarios, the proposed RBF-Coulomb GAN outperforms the baselines (lower $\mathcal{W}^{2,2}$ scores). RBF-Coulomb GAN is also less sensitive to learning-rate perturbations, compared to the trainable-discriminator counterparts. These results showcase the superiority of solving the discriminator PDE over stochastic-gradient optimization.

5 Conclusions

We showed that the Coulomb GAN discriminator is, in fact, the optimum to a special case of gradient-regularized WGAN. More precisely, the solution comes up as the Green's function to the Poisson PDE, when approximated using RBFs. Considering toy examples, we demonstrate the superior convergence of employing the closed-form RBF discriminator over the baseline methods. To derive the generalization to Plummer-like kernels for $d \leq n - 2$, one could consider higher-order gradient penalties, which give rise to PDEs involving iterated Laplacian operators. Exploring PDE solvers to implement the optimal discriminator in image learning scenario is a promising direction for research. Results from denoising diffusion models (Ho et al., 2020; Xiao et al., 2022) and score-matching networks (Song & Ermon, 2019, 2020; Song et al., 2021) could be leveraged to learn the optimal Coulomb GAN discriminator in a way that avoids issues with training such as learning-rate scheduling and mode collapse. One could also derive PDEs and associated WGAN costs that correspond to generalizations of Plummer kernel dimension d. The variational approach can also be used to derive PDEs that govern other GAN discriminators that do not conform to IPM minimization, such as those with relativistic discriminators (Jolicoeur-Martineau, 2019).

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.

Aronszajn, N., Creese, T., and Lipkin, L. *Polyharmonic Functions*. Oxford: Clarendon, 1983.

Bellemare, M. G., Danihelka, I., Dabney, W., Mohamed, S., Lakshminarayanan, B., Hoyer, S., and Munos, R. The Cramér distance as a solution to biased Wasserstein gradients. 2017.

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

Fasshauer, G. E. *Meshfree Approximation Methods with MATLAB*. World Scientific, 2007.

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`.

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.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.

Iske, A. Multiresolution methods in scattered data modelling. In *Lecture Notes in Computational Science and Engineering*, 2004.

Jolicoeur-Martineau, A. The relativistic discriminator: a key element missing from standard GAN. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

Kang, M., Shin, J., and Park, J. StudioGAN: A taxonomy and benchmark of GANs for image synthesis, 2022. URL `https://arxiv.org/abs/2206.09479`.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2019.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., and Aila, T. Training generative adversarial networks with limited data. In *Advances in Neural Information Processing Systems 33*, 2020.

Karras, T. et al. Alias-free generative adversarial networks, 2021.

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`.

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. 2017a.

Li, J., Madry, A., Peebles, J., and Schmidt, L. Towards understanding the dynamics of generative adversarial networks. *arXiv preprints, arXiv:1706.09884*, 2017b. URL `http://arxiv.org/abs/1706.09884`.

Li, Y., Swersky, K., and Zemel, R. Generative moment matching networks. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1718–1727, Jul 2015.

Mescheder, L., Geiger, A., and Nowozin, S. Which training methods for GANs do actually converge? In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 3481–3490, 2018.

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

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

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, 2019.

Song, Y. and Ermon, S. Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems*, 2020.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=PxTIG12RRHS`.

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

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=SkVqX0xCb`.

Xiao, Z., Kreis, K., and Vahdat, A. Tackling the generative learning trilemma with denoising diffusion GANs. In *Proceeding of the 10th International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=JprM0p-q0Co`.

A Proof of Theorem 3.1: Gradient-regularized WGANs

First, we recall a few useful results from the *Calculus of Variations*. Consider the functional optimization of a cost $\mathcal{L}$ defined as

$$\mathcal{L}\left(f(\boldsymbol{x}), \{\partial_i f(\boldsymbol{x})\}\right) = \int_{\mathcal{X}} \mathcal{F}\left(\boldsymbol{x}, f(\boldsymbol{x}), \{\partial_i f(\boldsymbol{x})\}\right) \, \mathrm{d}\boldsymbol{x}, \tag{8}$$

where $\partial_i f(\boldsymbol{x}) = \dfrac{\partial f}{\partial x_i}$ denotes the partial derivative of f with respect to the ith entry of $\boldsymbol{x}$. The Euler-Lagrange condition that the optimizer $f^*(\boldsymbol{x})$ of the cost $\mathcal{L}$ must satisfy is given by

$$\frac{\partial \mathcal{F}}{\partial f} - \sum_{i=1}^{N} \left[\frac{\partial}{\partial x_i} \left(\frac{\partial \mathcal{F}}{\partial(\partial_i f)} \right) \right] \Bigg|_{f=f^*(\boldsymbol{x})} = 0. \tag{9}$$

We can now derive the optimal discriminator in gradient-regularized WGAN. Recall that

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

where $\rho(\boldsymbol{x}) = p_g(\boldsymbol{x}) - p_d(\boldsymbol{x})$, and $\|\nabla D(\boldsymbol{x})\|_2^2 = \sum_i (\partial_i D)^2$ is the square of the ℓ_2 norm of the gradient vector. Differentiating the integrand in $\mathcal{L}_D$ gives

$$\frac{\partial \mathcal{F}}{\partial f} = \rho(\boldsymbol{x}) \quad \text{and} \quad \sum_{i=1}^{N} \left[\frac{\partial}{\partial x_i} \left(\frac{\partial \mathcal{F}}{\partial(\partial_i f)} \right) \right] = 2\lambda_d \sum_{i=1}^{N} \partial_i \left(\partial_i D(\boldsymbol{x}) \right) = 2\lambda_d \Delta D(\boldsymbol{x}).$$

Enforcing the Euler-Lagrange condition yields the PDE given in Equation (5)

$$\Delta D(\boldsymbol{x}) = \frac{\rho(\boldsymbol{x})}{2\lambda_d}, \; \forall \, \boldsymbol{x} \in \mathcal{X}.$$

The above Poisson PDE can be solved to obtain the optimal discriminator function. As noted by Evans (2010); Unterthiner et al. (2018), the family of Coulomb kernels $\psi_n(\boldsymbol{x})$ are Green's functions to the Laplace operator, *i.e.,* $\Delta\psi_n(\boldsymbol{x}) = \mathrm{C}_n \, \delta(\boldsymbol{x})$, where $\mathrm{C}_n = \frac{\Gamma\left(\frac{n}{2}+1\right)}{n(n-2)\pi^{\frac{n}{2}}}$, where in turn $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} \, \mathrm{d}t$ denotes the Gamma function. Convolving both sides of the Poisson PDE with ψ_n gives the optimal discriminator:

$$\begin{aligned}
\Delta \left(D * \psi_n\right)(\boldsymbol{x}) = \left(D * \Delta\psi_n\right)(\boldsymbol{x}) = \left(D * \mathrm{C}_n \, \delta\right)(\boldsymbol{x}) \Big|_{D=D^*} &= \frac{1}{2\lambda_d} \left(\rho * \psi_n\right)(\boldsymbol{x}) \\
\Rightarrow D^*(\boldsymbol{x}) &= \frac{1}{2\mathrm{C}_n \lambda_d} \left(\rho * \psi_n\right)(\boldsymbol{x}), \;\; \forall \, \boldsymbol{x} \in \mathcal{X}.
\end{aligned}$$

This completes the proof of Theorem 3.1.

B Proof of Lemma 3.2: Sample Estimate of the Optimal Discriminator

Consider the closed-form optimal discriminator given in Equation (6):

$$D_p^*(\boldsymbol{x}) = \frac{1}{2\lambda_d} \left(\rho * \psi_n\right)(\boldsymbol{x}) = \frac{1}{2\lambda_d} \left((p_g - p_d) * \psi_n\right)(\boldsymbol{x}),$$

where λ_d accounts for the effect of C_n. Without loss of generality, we assume that $n > 2$. From the definition of the convolution, we have

$$\begin{aligned}
D_p^*(\boldsymbol{x}) &= \frac{1}{2\lambda_d} \int_{\mathcal{X}} \left(p_g(\boldsymbol{y}) - p_d(\boldsymbol{y})\right) \|\boldsymbol{x} - \boldsymbol{y}\|^{2-n} \, \mathrm{d}\boldsymbol{y} \\
&= \frac{1}{2\lambda_d} \Big(\mathbb{E}_{\boldsymbol{y}\sim p_g}[\|\boldsymbol{x} - \boldsymbol{y}\|^{2-n}] - \mathbb{E}_{\boldsymbol{y}\sim p_d}[\|\boldsymbol{x} - \boldsymbol{y}\|^{2-n}] \Big).
\end{aligned}$$

As in the case of Coulomb GAN (Unterthiner et al., 2018), the expectations can now be replaced with N-sample estimates as follows:

$$\tilde{D}_p^*(\boldsymbol{x}) = \frac{1}{2N_g\lambda_d}\sum_{\boldsymbol{g}_j\sim p_g}\|\boldsymbol{x}-\boldsymbol{g}_j\|^{2-n} - \frac{1}{2N_d\lambda_d}\sum_{\boldsymbol{d}_i\sim p_d}\|\boldsymbol{x}-\boldsymbol{d}_i\|^{2-n},$$

where $\tilde{D}_p^*$ is essentially a radial basis function expansion.

A similar analysis could be carried out for $n = 1,\ 2$, and the corresponding discriminators are:

$$\tilde{D}_p^*(\boldsymbol{x}) = \begin{cases} \dfrac{1}{2N_g\lambda_d}\displaystyle\sum_{g_j\sim p_g}|x-g_i| - \dfrac{1}{2N_d\lambda_d}\displaystyle\sum_{d_i\sim p_d}|x-d_i|, & \text{for } n=1, \\ \dfrac{1}{2N_d\lambda_d}\displaystyle\sum_{\boldsymbol{d}_i\sim p_d}\ln(\|\boldsymbol{x}-\boldsymbol{d}_i\|) - \dfrac{1}{2N_g\lambda_d}\displaystyle\sum_{\boldsymbol{g}_j\sim p_g}\ln(\|\boldsymbol{x}-\boldsymbol{g}_j\|), & \text{for } n=2. \end{cases}$$

The generic form for $\tilde{D}_p^*$ is given by

$$\tilde{D}_p^*(\boldsymbol{x}) = \frac{1}{2N_g\lambda_d}\sum_{\boldsymbol{g}_j\sim p_g}\psi_n(\boldsymbol{x}-\boldsymbol{g}_j) - \frac{1}{2N_d\lambda_d}\sum_{\boldsymbol{d}_i\sim p_d}\psi_n(\boldsymbol{x}-\boldsymbol{d}_j),$$

where ψ_n is as defined in Equation (3).

C Network Specifications and Training Parameters

Data: We consider unimodal learning on a 2-D Gaussian $\mathcal{N}(\mathbf{3.5}, 1.25\mathbb{I})$, where $\mathbf{3.5}$ denotes a vector with both values equal to 3.5, and $\mathbb{I}$ is the identity matrix. For the multimodal learning task, we consider an eight-component Gaussian mixture model (GMM) with component means selected from $[0,1]^2$, each having standard deviation 0.02.

Network: We consider the affine transformation of the noise $\boldsymbol{z}$, given by $\boldsymbol{x} = M\boldsymbol{z} + b$ to learn 2-D Gaussian data. The discriminator for RBF-Coulomb GAN is an RBF network whose weights are computed out-of-the-loop, thereby obviating the need for network-based optimization of the discriminator. On the GMM task, the generator consists of three fully-connected layers of 10, 10 and 5 nodes, followed by an output layer of 2 nodes. The activation used is Leaky ReLU. The discriminator in the baselines consists of three fully-connected layers with Leaky ReLU activation and 20, 10, and 1 node(s). The discriminator is identical for both the Gaussian and GMM learning tasks.

Training Parameters: The networks are trained using the Adam optimizer (Kingma & Ba, 2015) with a learning rate of 0.02 for the generator and 0.075 for the discriminator. The batch size is set to 500. The models are evaluated using the Wasserstein-2 distance, estimated using the *Python optimal transport (POT)* repository (Flamary et al., 2021).

Source Code: The Source code for *RBF-Coulomb GANs* is developed in TensorFlow 2.0 (Abadi et al., 2016) and will be released publicly on GitHub with the camera-ready version of this manuscript.