The WROF Problem and WGANs

Cameron Davies

Department of Mathematics
University of Toronto

December 4, 2023

AN OPTIMAL TRANSPORT ANALOGUE OF THE RUDIN OSHER FATEMI MODEL AND ITS CORRESPONDING MULTISCALE THEORY*

TRISTAN MILNE† AND ADRIAN NACHMAN†‡

Abstract. We develop a theory for image restoration with a learned regularizer that is analogous to that of Meyer's characterization of solutions of the classical variational method of Rudin-Osher-Fatemi (ROF). The learned regularizer we use is a Kantorovich potential for an optimal transport problem of mapping a distribution of noisy images onto clean ones, as first proposed by Lunz, Öktem and Schönlieb. We show that the effect of their restoration method on the distribution of the images is an explicit Euler discretization of a gradient flow on probability space, while our variational problem, dubbed Wasserstein ROF (WROF), is the corresponding implicit discretization. We obtain our geometric characterisation of the solution in this setting by first proving a more general convex analysis theorem for variational problems with solutions characterised by projections. We then use optimal transport arguments to obtain our WROF theorem from this general result, as well as a decomposition of a transport map into large scale "features" and small scale "details", where scale refers to the magnitude of the transport distance. Further, we leverage our theory to analyze two algorithms which iterate WROF. We refer to these as iterative regularization and multiscale transport. For the former we prove convergence to the clean data. For the latter we produce successive approximations to the target distribution that match it up to finer and finer scales. These algorithms are in complete analogy to well-known effective methods based on ROF for iterative denoising, respectively hierarchical image decomposition. We also obtain an analogue of the Tadmor Nezzar Vese energy identity which decomposes the Wasserstein 2 distance between two measures into a sum of non-negative terms that correspond to transport costs at different scales.

Background: GANs

- **Goal:** generate fake data μ which is hard to distinguish from real data ν. Originally proposed by Goodfellow et. al. in [6].

Background: GANs

- **Goal:** generate fake data μ which is hard to distinguish from real data ν. Originally proposed by Goodfellow et. al. in [6].
- Generative adversarial networks consist of a generator neural network $G_\theta : \mathbb{R}^m \to \mathbb{R}^d$ and a discriminator/critic neural network $C_{\tilde{\theta}} : \mathbb{R}^d \to [0,1]$, estimating the probability that a datum is fake.
- Networks are trained iteratively to attain:

$$\min_{\theta} \max_{\tilde{\theta}} \left[\int \log C_{\tilde{\theta}} d\nu + \int \log(1 - C_{\tilde{\theta}} \circ G_\theta) d\xi \right],$$

where ξ is Gaussian.

Background: GANs

- **Goal:** generate fake data μ which is hard to distinguish from real data ν. Originally proposed by Goodfellow et. al. in [6].
- Generative adversarial networks consist of a generator neural network $G_\theta : \mathbb{R}^m \to \mathbb{R}^d$ and a discriminator/critic neural network $C_{\tilde{\theta}} : \mathbb{R}^d \to [0,1]$, estimating the probability that a datum is fake.
- Networks are trained iteratively to attain:

$$\min_{\theta} \max_{\tilde{\theta}} \left[\int \log C_{\tilde{\theta}} d\nu + \int \log(1 - C_{\tilde{\theta}} \circ G_\theta) d\xi \right],$$

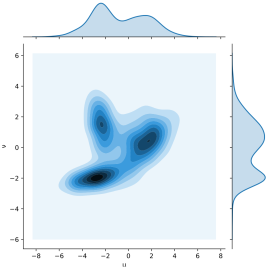
  where ξ is Gaussian.
- Critic uses the Jensen-Shannon divergence to try to distinguish ν and $G_{\theta\#}\xi$.

Background: Wasserstein Distance

- Given probability measures μ, $\nu \in \mathcal{P}(\mathbb{R}^d)$ and $p > 1$, define

$$d_p(\mu,\nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \left[\int_{\mathbb{R}^n} |x-y|^p d\gamma(x,y) \right]^{1/p}.$$

- $\Gamma(\mu,\nu)$ is the set of joint distributions with marginals μ and ν.



Background: WGANs and TTC

- Wasserstein Generative Adversarial Networks (WGANs), introduced in [4], use the Wasserstein d_1 distance to distinguish ν and $G_{\theta\#}\xi$.
- Kantorovich-Rubenstein Theorem: for $\Omega \subseteq \mathbb{R}^d$

$$d_1(\mu, \nu) = \sup_{u \in 1\text{-Lip}(\Omega)} \left[\int_\Omega u d\mu - \int_\Omega u d\nu \right],$$

so WGANs train a critic $u_{\tilde{\theta}}$ to attain the supremum (Kantorovich potential).

Background: WGANs and TTC

- Wasserstein Generative Adversarial Networks (WGANs), introduced in [4], use the Wasserstein d_1 distance to distinguish ν and $G_{\theta\#}\xi$.
- Kantorovich-Rubenstein Theorem: for $\Omega \subseteq \mathbb{R}^d$

$$d_1(\mu,\nu) = \sup_{u \in 1\text{-Lip}(\Omega)} \left[\int_\Omega u d\mu - \int_\Omega u d\nu \right],$$

  so WGANs train a critic $u_{\tilde{\theta}}$ to attain the supremum (Kantorovich potential).
- Milne's "Trust the Critics" Algorithm in [7, 9]: introduce a gradient penalty to account for the Lipschitz condition. Then replace the generator with a procedure for editing the data by gradient descent using trained critics.

Background: Example of TTC

https://raw.githubusercontent.com/tmilne5/Trust-the-Critics-2/main/TTC_gifs/photo-2-monet.gif

Iterative Regularization and the WROF Problem

- In relevant cases, one step of the TTC algorithm is equivalent to the following "Wasserstein Rudin-Osher-Fatemi" problem proposed in [7, 8] based on [10].
- Given $\rho_0^\tau = \mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ and $\tau > 0$, find

$$\rho_{k+1}^\tau \in \operatorname*{argmin}_{\rho \in \mathcal{P}(\mathbb{R}^d)} \left[\frac{1}{2\tau} d_2^2(\rho_k^\tau, \rho) + d_1(\rho, \nu) \right].$$

Iterative Regularization and the WROF Problem

- In relevant cases, one step of the TTC algorithm is equivalent to the following "Wasserstein Rudin-Osher-Fatemi" problem proposed in [7, 8] based on [10].
- Given $\rho_0^\tau = \mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ and $\tau > 0$, find

$$\rho_{k+1}^\tau \in \underset{\rho \in \mathcal{P}(\mathbb{R}^d)}{\operatorname{argmin}} \left[\frac{1}{2\tau} d_2^2(\rho_k^\tau, \rho) + d_1(\rho, \nu) \right].$$

- Issues with d_1 : not strictly convex, non-unique Kantorovich potential, not covered by minimizing movement scheme theory of [3].

Iterative Regularization and the WROF Problem

- In relevant cases, one step of the TTC algorithm is equivalent to the following "Wasserstein Rudin-Osher-Fatemi" problem proposed in [7, 8] based on [10].
- Given $\rho_0^\tau = \mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ and $\tau > 0$, find

$$\rho_{k+1}^\tau \in \operatorname*{argmin}_{\rho \in \mathcal{P}(\mathbb{R}^d)} \left[\frac{1}{2\tau} d_2^2(\rho_k^\tau, \rho) + d_1(\rho, \nu) \right].$$

- Issues with d_1 : not strictly convex, non-unique Kantorovich potential, not covered by minimizing movement scheme theory of [3].
- **Idea:** see what happens when we use the modified scheme

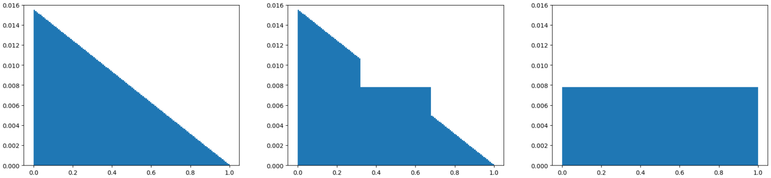
$$\rho_{k+1}^\tau \in \operatorname*{argmin}_{\rho \in \mathcal{P}(\mathbb{R}^d)} \left[\frac{1}{2\tau} d_2^2(\rho_k^\tau, \rho) + d_{1+\tau}^{1+\tau}(\rho, \nu) \right].$$

Numerical Experiments

- **Unsurprising:** If τ is too small, $\rho_k = \mu$ for all $k \geq 1$. If τ is too large, $\rho_k = \nu$ for all $k \geq 1$.

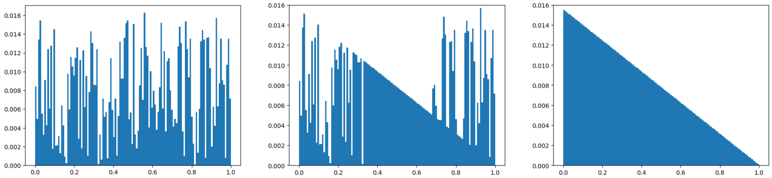
Numerical Experiments

- **Unsurprising:** If τ is too small, $\rho_k = \mu$ for all $k \geq 1$. If τ is too large, $\rho_k = \nu$ for all $k \geq 1$.
- **Surprising:** For τ in the middle, you get what appears to be a steady state after the first step:



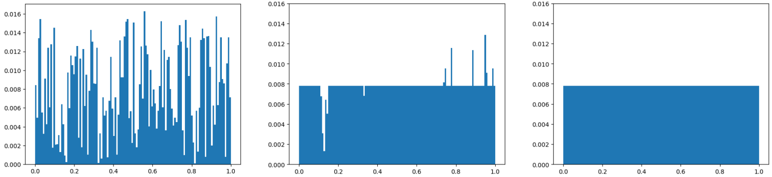
Numerical Experiments

- **Unsurprising:** If τ is too small, $\rho_k = \mu$ for all $k \geq 1$. If τ is too large, $\rho_k = \nu$ for all $k \geq 1$.
- **Surprising:** For τ in the middle, you get what appears to be a steady state after the first step:



Numerical Experiments

- **Unsurprising:** If τ is too small, $\rho_k = \mu$ for all $k \geq 1$. If τ is too large, $\rho_k = \nu$ for all $k \geq 1$.
- **Surprising:** For τ in the middle, you get what appears to be a steady state after the first step:



Toy Model Explanation

- **Example:** Consider the space $\{0, r\}$, measures $\mu = \delta_0$, and $\nu = \delta_r$, and represent $\rho \in \mathcal{P}(\{0, r\})$ as $\rho = \rho^0 \delta_0 + (1 - \rho^0)\delta_r$, for $\rho_0 \in [0, 1]$.

Toy Model Explanation

- **Example:** Consider the space $\{0, r\}$, measures $\mu = \delta_0$, and $\nu = \delta_r$, and represent $\rho \in \mathcal{P}(\{0, r\})$ as $\rho = \rho^0\delta_0 + (1 - \rho^0)\delta_r$, for $\rho_0 \in [0, 1]$.
- Then for $\alpha \in \{1, 1 + \tau\}$,

$$\frac{1}{2\tau} d_2^2(\mu, \rho) + d_\alpha^\alpha(\rho, \nu) = (r^\alpha - \frac{r^2}{2\tau})\rho^0 + \frac{r^2}{2\tau}.$$

- If $r^{2-\alpha} < 2\tau$, minimized at $\rho^0 = 0$. If $r^{2-\alpha} > 2\tau$, minimized at $\rho^0 = 1$. If $r^{2-\alpha} = 2\tau$, constant for $\rho^0 \in [0, 1]$.

Toy Model Explanation

- **Example:** Consider the space $\{0, r\}$, measures $\mu = \delta_0$, and $\nu = \delta_r$, and represent $\rho \in \mathcal{P}(\{0, r\})$ as $\rho = \rho^0\delta_0 + (1 - \rho^0)\delta_r$, for $\rho_0 \in [0, 1]$.
- Then for $\alpha \in \{1, 1 + \tau\}$,

$$\frac{1}{2\tau} d_2^2(\mu, \rho) + d_\alpha^\alpha(\rho, \nu) = (r^\alpha - \frac{r^2}{2\tau})\rho^0 + \frac{r^2}{2\tau}.$$

- If $r^{2-\alpha} < 2\tau$, minimized at $\rho^0 = 0$. If $r^{2-\alpha} > 2\tau$, minimized at $\rho^0 = 1$. If $r^{2-\alpha} = 2\tau$, constant for $\rho^0 \in [0, 1]$.
- **Upshot:** Movement is instant on small scales and forbidden on large scales

Toy Model Explanation

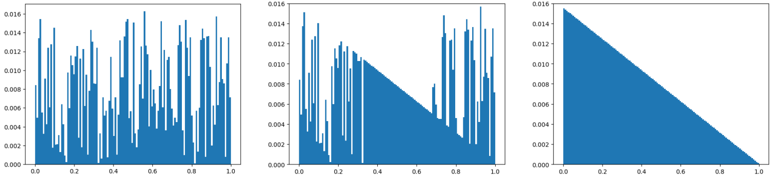
- **Example:** Consider the space $\{0, r\}$, measures $\mu = \delta_0$, and $\nu = \delta_r$, and represent $\rho \in \mathcal{P}(\{0, r\})$ as $\rho = \rho^0\delta_0 + (1 - \rho^0)\delta_r$, for $\rho_0 \in [0, 1]$.
- Then for $\alpha \in \{1, 1 + \tau\}$,

$$\frac{1}{2\tau} d_2^2(\mu, \rho) + d_\alpha^\alpha(\rho, \nu) = (r^\alpha - \frac{r^2}{2\tau})\rho^0 + \frac{r^2}{2\tau}.$$

- If $r^{2-\alpha} < 2\tau$, minimized at $\rho^0 = 0$. If $r^{2-\alpha} > 2\tau$, minimized at $\rho^0 = 1$. If $r^{2-\alpha} = 2\tau$, constant for $\rho^0 \in [0, 1]$.
- **Upshot:** Movement is instant on small scales and forbidden on large scales
- **Role of α:** Higher α makes jumps of distance > 1 easier, but jumps of distance < 1 harder.

Numerical Experiments

- **Unsurprising:** If τ is too small, $\rho_k = \mu$ for all $k \geq 1$. If τ is too large, $\rho_k = \nu$ for all $k \geq 1$.
- **Surprising:** For τ in the middle, you get what appears to be a steady state after the first step:



Future Directions

- The algorithm seems to work fine at small scales, so it might be worthwhile to see if a multiscale approach as in [8] works.
- A better algorithm for computing the WROF problem (e.g. in [1, 5]) could be beneficial, especially since a curse of dimensionality seems likely in analogy to [2].
- A different style of numerics – e.g. a particle simulation on a very large discrete grid could possibly act more like the continuous case (where convergence is known under light assumptions due to [8]).

Thank you!

Any questions?

References

[1] Jason M. Altschuler and Enric Boix-Adserà, *Wasserstein barycenters can be computed in polynomial time in fixed dimension*, J. Mach. Learn. Res. **22** (2021), Paper No. 44, 19. MR4253737

[2] ______, *Wasserstein barycenters are NP-hard to compute*, SIAM J. Math. Data Sci. **4** (2022), no. 1, 179–203. MR4378594

[3] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré, *Gradient flows in metric spaces and in the space of probability measures*, Second, Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, 2008. MR2401600

[4] Martin Arjovsky, Soumith Chintala, and Léon Bottou, *Wasserstein generative adversarial networks*, Proceedings of the 34th international conference on machine learning, 201706, pp. 214–223.

[5] Jiaojiao Fan, Qinsheng Zhang, Amirhossein Taghvaei, and Yongxin Chen, *Variational Wasserstein gradient flow*, Proceedings of the 39th international conference on machine learning, 202217, pp. 6185–6215.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, *Generative adversarial nets*, Advances in neural information processing systems **27** (2014).

[7] Tristan Milne, *Optimal transport, congested transport, and wasserstein generative adversarial networks*, PhD thesis, Toronto, Ontario, 2022. Available at `https://hdl.handle.net/1807/125470`.

[8] Tristan Milne and Adrian Nachman, *An optimal transport analogue of the rudin osher fatemi model and its corresponding multiscale theory*, 2023.

[9] Tristan Milne, Étienne Bilocq, and Adrian Nachman, *Trust the critics: Generatorless and multipurpose wgans with initial convergence guarantees*, 2021.

[10] Leonid I. Rudin, Stanley Osher, and Emad Fatemi, *Nonlinear total variation based noise removal algorithms*, 1992, pp. 259–268. Experimental mathematics: computational issues in nonlinear science (Los Alamos, NM, 1991). MR3363401