ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414895

MSR-GAN: MULTI-SEGMENT RECONSTRUCTION VIA ADVERSARIAL LEARNING

Mona Zehni, Zhizhen Zhao

Department of ECE and CSL, University of Illinois at Urbana-Champaign

ABSTRACT

Multi-segment reconstruction (MSR) is the problem of estimating a signal given noisy partial observations. Here each observation corresponds to a randomly located segment of the signal. While previous works address this problem using template or moment-matching, in this paper we address MSR from an unsupervised adversarial learning standpoint, named MSR-GAN. We formulate MSR as a *distribution* matching problem where the goal is to recover the signal and the probability distribution of the segments such that the distribution of the generated measurements following a known forward model is close to the real observations. This is achieved once a min-max optimization involving a generator-discriminator pair is solved. MSR-GAN is mainly inspired by CryoGAN [1]. However, in MSR-GAN we no longer assume the probability distribution of the latent variables, i.e. segment locations, is given and seek to recover it alongside the unknown signal. For this purpose, we show that the loss at the generator side originally is non-differentiable with respect to the segment distribution. Thus, we propose to approximate it using Gumbel-Softmax reparametrization trick. Our proposed solution is generalizable to a wide range of inverse problems. Our simulation results and comparison with various baselines verify the potential of our approach in different settings.

Index Terms— Multi-segment reconstruction, adversarial learning, unsupervised learning, Gumbel-Softmax approximation, categorical distribution.

1. INTRODUCTION

The problem of recovering a signal from a set of noisy partial observations appear in a wide range of applications including genomic sequence assembly [2], puzzle solving[3], tomographic reconstruction [4] and cryo-electron microscopy (Cryo-EM) [5, 6], to name a few. In this paper, we focus on multi-segment reconstruction (MSR) [7], where the unknown is a 1D sequence and the measurements are noisy randomly located partial observations (segments) of this sequence. A schematic illustration of MSR is provided in Fig. 1. MSR is a general form of multi-reference alignment (MRA) [8] problem in which the measurements are noisy randomly shifted versions of the signal. While in MRA the length of each measurement is the same as the signal, in MSR the measurements can be shorter.

Current efforts devoted to MSR is studied in two broad categories, 1) alignment-based, 2) alignment-free. In one form of alignment-based methods, the segment location corresponding to each observation is estimated. Then, the observations are aligned accordingly and averaged. While these methods have low computational and sample complexity, low signal-to-noise ratio (SNR) of the observations adversely affect their performance. Examples of alignment-based methods applied to MRA and tomographic reconstruction are found in [9, 10]. In other forms of alignment based methods, the segment locations and the 1D sequence are jointly updated using alternating steps. An example would be the maximum likelihood formulation of MSR, solved using expectation-maximization (EM). Despite the robustness of EM to different noise regimes, it suffers from high computational complexity. This is due to the complexity of the E-step, requiring a whole pass through the measurements at every iteration. This is significantly time-consuming, especially in the presence of large number of observations.

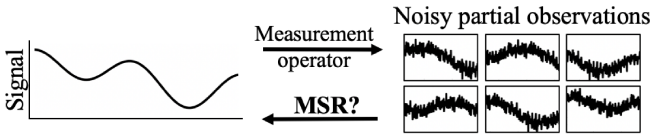

Fig. 1. Multi-segment reconstruction (MSR) problem.

Alignment-free solutions specifically designed for MRA sidestep the estimation of the random shifts by introducing a set of invariant features. These features constitute the moments of the signal and are estimated from the measurements. The signal is then estimated from the features via an optimization-based framework [8, 11], tensor decomposition [12, 13] using Jennrich's algorithm [14] or spectral decomposition [15, 16]. As these works are specialized for MRA, they do not address the challenges associated with MSR, such as observing only shorter segments of the signal. In [7], we showed how for MSR, we can estimate the invariant features from the measurements and how the recovery of the signal is tied to the segment length. Compared to alignment-based solutions, in alignment-free methods, we only have one pass through the measurements to estimate the features, thus computationally more efficient. The estimated features then serve as a compact representation of the measurements which are functions of the unknown signal and segment location distribution.

In this paper, we propose an alignment-free adversarial learning based method for MSR. Our goal is to find the unknown 1D signal and the distribution of the segment locations such that the measurements generated from the estimated signal match the real measurements in a distribution sense. Therefore, we train a generator discriminator pair, where the discriminator tries to distinguish between the measurements output by the generator and the real ones. Our approach is inspired by CryoGAN [1] in which the goal is to reconstruct a 3D structure given 2D noisy projection images from unknown projection views. Unlike CryoGAN, we assume the distribution of the latent variables, i.e. the segment locations in MSR, is unknown and we seek to recover it alongside the signal. For this purpose, we modify the loss at the generator side using Gumbel-Softmax approximation of categorical distribution, to accommodate gradient-based updates of the segment location distribution. Our simulation results and comparison with several baselines confirm the feasibility of our approach in various segment length and noise regimes. Our code is available at https://github.com/MonaZI/MSR-GAN.

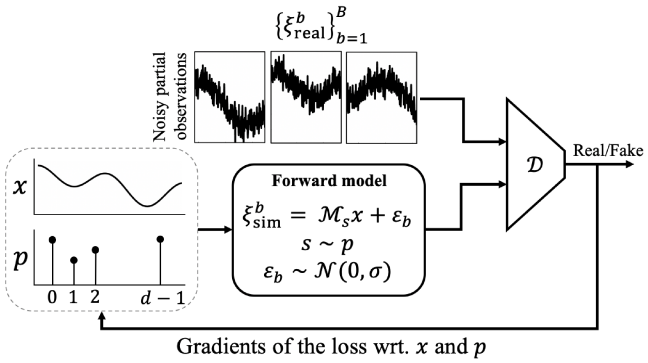

Fig. 2. An illustration of MSR-GAN pipeline.

2. SYSTEM MODEL

We consider the following observation model,

$$\xi_j = \mathcal{M}_{s_j} x + \varepsilon_j, \quad j \in \{1, 2, ..., N\} \tag{1}$$

where $x \in \mathbb{R}^d$ is the underlying signal and $\xi_j \in \mathbb{R}^m$, $m \leq d$ is the j-th observation. We often refer to m as the segment length. The cyclic masking operator $\mathcal{M}_s$ captures m consecutive entries of x starting from index s. In other words, $\mathcal{M}_s : \mathbb{R}^d \to \mathbb{R}^m$ and $(\mathcal{M}_s x)[n] = x[n + s \bmod d]$. We also assume the segment location $s \in \{0, 1, ..., d-1\}$ to be unknown and randomly drawn from a categorical distribution with p as its probability mass function (PMF) where $P\{s = s_j\} = p[s_j]$. In addition, the randomly located segment of the signal is contaminated by additive white Gaussian noise ε_j with zero mean and covariance $\sigma^2 I_m$ (I_m denoting the identity matrix with size $m \times m$). Our goal here is to recover x and p given the noisy partial observations $\{\xi_j\}_{j=1}^N$.

Note that the distribution of the observations depends on both the signal x and the distribution of the segment locations p. Thus, it is possible to estimate x and p by matching the distribution of the observations generated by x and p following (1) to the real measurements.

3. METHOD

We use an unsupervised adversarial learning approach to solve MSR. Our method is unsupervised as it only relies on the given observations and does not use large paired datasets for training. Similar to [1], our method aims to find x and p such that the distribution of the partial noisy measurements generated from (1) matches the real measurements $\{\xi^j_{\text{real}}\}_{j=1}^N$. To this end, we use a generative adversarial network (GAN) [17]. Unlike common GAN models, we use the known forward model in (1) to map the signal and segment distribution to the measurements. Thus, the generator acts upon x and p and simulates noisy measurements $\{\xi^j_{\text{sim}}\}_{j=1}^M$. The discriminator's task is then to distinguish between the real and fake measurements from the generator. An illustration of MSR-GAN is provided in Fig. 2. Here we use Wasserstein GAN [18] with gradient penalty (WGAN-GP) [19], to benefit from its favorable convergence behaviour. In WGAN, the output of the discriminator is a score, where the more the input resembles ξ_{real}, the higher the score. The min-max formulation of the problem is:

$$\widehat{x}, \widehat{p} = \arg\min_{x,p} \max_{\phi} \mathcal{L}(\phi, x, p) \tag{2}$$

$$\mathcal{L}(\phi, x, p) = \sum_{b=1}^{B} \mathcal{D}_\phi(\xi^b_{\text{real}}) - \mathcal{D}_\phi(\xi^b_{\text{sim}}) - \lambda\, \text{GP}(\xi^b_{\text{int}}) \tag{3}$$

Algorithm 1 MSR-GAN

Require: α_ϕ, α_x, α_p: learning rates for the discriminator, the image and projection angle distribution. λ: gradient penalty weight. n_{disc}: the number of iterations of the discriminator (critic) per generator iteration.

Require: Initialize x randomly and p with a uniform distribution, i.e. $p^0[s] = 1/d$.

Output: Estimates I and p given $\{\xi^j_{\text{real}}\}_{j=1}^N$.

1: **while** not converged **do**
2: **for** $t = 0, ..., n_{\text{disc}-1}$ **do**
3: Sample a batch from real data, $\{\xi^b_{\text{real}}\}_{b=1}^B$
4: Sample a batch of simulated measurements using estimated signal and PMF, i.e. $\{\xi^b_{\text{sim}}\}_{b=1}^B$ where $\xi^b_{\text{sim}} = \mathcal{M}_s x + \varepsilon_b$, $\varepsilon_b \sim \mathcal{N}(0, \sigma I_m)$
5: Generate interpolated samples $\{\xi^b_{\text{int}}\}_{b=1}^B$, $\xi^b_{\text{int}} = \alpha\, \xi^b_{\text{real}} + (1-\alpha)\, \xi^b_{\text{sim}}$ with $\alpha \sim \text{Unif}(0, 1)$
6: Update the discriminator using gradient ascent steps with,

$$\nabla_\phi \mathcal{L}_D(\phi) = \nabla_\phi \left(\sum_{b=1}^{B} \mathcal{D}_\phi(\xi^b_{\text{real}}) - \mathcal{D}_\phi(\xi^b_{\text{sim}}) + \lambda \text{GP}(\xi^b_{\text{int}}) \right)$$

7: **end for**
8: Sample a batch of $\{q_{i,b}\}_{b=1}^B$ using (8)
9: Update x and p using gradient descent steps with the following gradients,

$$\nabla_{x,p} \mathcal{L}_G(x, p) = \nabla_{x,p} \left(-\sum_{b=1}^{B} \sum_{s=0}^{d-1} q_{i,b} \mathcal{D}_\phi(\mathcal{M}_s x + \varepsilon_b) \right)$$

10: **end while**

$$\text{GP}(\xi^b_{\text{int}}) = \left(\|\nabla_\xi \mathcal{D}_\phi(\xi^b_{\text{int}})\| - 1 \right)^2 \tag{4}$$

where $\mathcal{L}$ denotes the loss which is a function of the discriminator's parameters ϕ, the signal and the PMF. Also, B is the batch size, $\mathcal{D}_\phi$ denotes the discriminator parameterized by ϕ and $\xi_{\text{sim}} = \mathcal{M}_s x + \varepsilon$, $s \sim p$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_m)$. The weight of the gradient penalty term (GP) is λ and ξ_{int} is a sample generated by linear interpolation between a real and simulated measurement, i.e. $\xi_{\text{int}} = \alpha\, \xi_{\text{real}} + (1 - \alpha)\, \xi_{\text{sim}}$, $\alpha \sim$ Unif(0, 1). To solve the min-max optimization in (2), following common practice, we take alternating steps to update the discriminator's parameters ϕ and the generator, i.e. x and p, using their gradients.

To update p, we need to take gradients of (3) with respect to p. However, this loss function is related to p through a sampling operator which is non-differentiable (we are sampling the segment locations based on the p distribution). This would be problematic at the generator update steps. Therefore, it is crucial to devise a way to have a meaningful gradient with respect to p. First, let us take a closer look at the loss function that is minimized at the generator side:

$$\mathcal{L}_G(x, p) = -\sum_{b=1}^{B} \mathcal{D}_\phi(\mathcal{M}_{s_b} x + \varepsilon_b),\ s \sim p,\ \varepsilon \sim \mathcal{N}(0, \sigma^2 I_m) \tag{5}$$

$$= -\sum_{b=1}^{B} \sum_{s=0}^{d-1} \delta(s - s_b) \mathcal{D}_\phi(\mathcal{M}_s x + \varepsilon_b) \tag{6}$$

where δ is the Kronecker delta and $\delta(s - s_b)$ denotes the one-hot representation of a sample drawn from a categorical distribution with PMF p. Jang et al. in [20] proposed a Gumbel-Softmax reparametrization trick to approximate samples from a categorical distribution with a differentiable function. We use this idea and

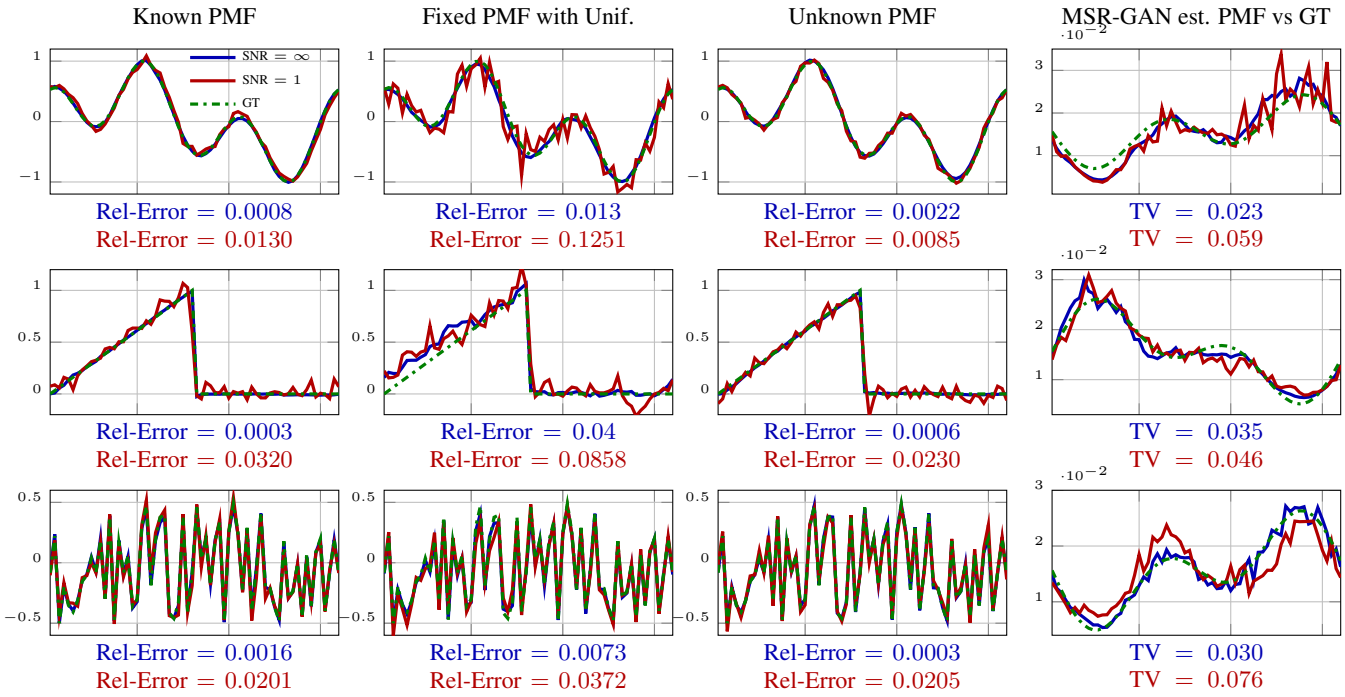

Fig. 3. Comparison between MSR-GAN in different noise regimes for 1) known PMF (first column), 2) unknown PMF but fixed with uniform distribution during training (second column), 3) unkown PMF and recovered during training (third column). The last column plots the ground truth PMF (green dashed curve) alongside the estimated PMFs from MSR-GAN (the same experiment as the third column) in blue and red. Each row corresponds to different signals and PMFs. The relative error of the reconstruction for SNR $= \infty$ and SNR $= 1$ is written in blue (SNR $= \infty$) and red (SNR $= 1$) underneath each subplot. For all experiments in this figure we are using the same architecture for the discriminator with $\ell = 100$ and the number of measurements is $N = 5 \times 10^4$.

replace $\delta(s - s_b)$, $s_b \sim p$ with a sample from the Gumbel-Softmax distribution, i.e.

$$\mathcal{L}_G(x, p) \approx \sum_{b=1}^{B} \sum_{s=0}^{d-1} q_{s,b}\, \mathcal{D}_\phi(\mathcal{M}_s x + \varepsilon_b) \tag{7}$$

where

$$q_{s,b} = \frac{\exp\left((g_{b,s} + \log(p[s]))/\tau\right)}{\sum_{i=0}^{d-1} \exp\left((g_{b,i} + \log(p[i]))/\tau\right)},\ g_{b,s} \sim \text{Gumbel}(0,1). \tag{8}$$

Note that (8) is a continuous approximation of the $\arg\max$ function, τ is the softmax temperature factor and $q_{s,b} \to \delta(s - \arg\max_s (g_{b,s} + \log p[s]))$ as $\tau \to 0$. Note that drawing samples from $\arg\max_s (g_{b,s} + \log p[s])$, $g_{b,s} \sim \text{Gumbel}(0,1)$ is an efficient way of sampling from p distribution [20]. Furthermore, to obtain samples from the Gumbel distribution [21], it suffices to transform samples from a uniform distribution using $g = -\log(-\log(u))$, $u \sim \text{Unif}(0,1)$.

4. IMPLEMENTATION DETAILS

We present the pseudo-code for MSR-GAN in Alg. 1. In all our experiments, we use a batch-size of $B = 200$ and keep the number of real measurements as $N = 3 \times 10^4$ unless otherwise mentioned. We have three separate learning rates for the discriminator, the signal and the PMF denoted by α_ϕ, α_x and α_p, while in most experimental settings we keep $\alpha_x = \alpha_p$. We reduce the learning rates by a factor of 0.9, with different schedules for different learning rates. We use `SGD` [22] as the optimizer for the discriminator and the signal x with a momentum of 0.9. We also update p using gradient descent steps after normalizing the corresponding gradients. We clip the gradients of the discriminator to have norm 1. Similar to common practice, we train the discriminator $n_{\text{disc}} = 4$ times per updates of x and p. To have stabilized updates with respect to p, we choose $\tau = 0.5$ in our experiments. We also use spectral normalization to stabilize the training [23].

Our architecture of the discriminator consists of three fully connected (FC) layers with ℓ, $\ell/2$, and 1 output sizes, where ℓ is determined accordingly for different experiments. We use `ReLU` [24] for the non-linear activations between the FC layers. We initialize the layers with weights drawn from normal distribution with mean zero and 0.01 standard deviation. We train MSR-GAN for $30,000$ and $50,000$ iterations for high and low SNR regimes, respectively. To enforce p to have non-negative values while adding up to one, we set it to be the output of a `Softmax` layer. Our implementation is in PyTorch and runs on single GPU.

5. NUMERICAL RESULTS

In this section we first provide details on our evaluation metrics and baselines. Next, we discuss our results.

Evaluation metrics and baselines: The SNR of the observations is defined as the variance of the clean measurements divided by the variance of the noise. As the signal and PMF are reconstructed up to a random global shift, we align the reconstructions before comparing them to the ground truths. We use relative error (rel-error) between the aligned estimated signal $\widehat{x}$ and the ground truth x as

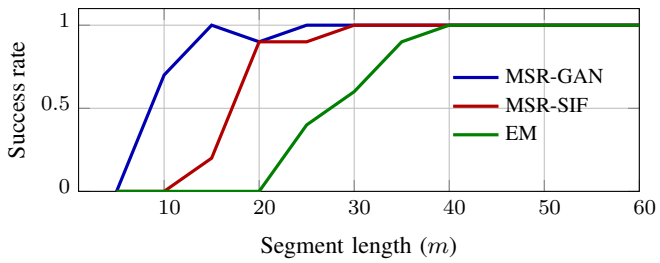

Fig. 4. Effect of segment length on the success rate of 1) MSR-GAN (blue curve), 2) MSR-SIF (red curve), 3) EM (green curve). In this experiment the signal length $d = 60$, the signal is generated randomly and $\sigma = 0.01$. The success rate is computed based on 10 random initializations for each segment length value. All three methods are initialized with the same random x and p.

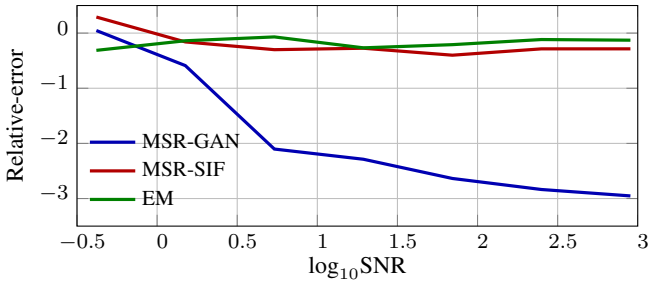

Fig. 5. Comparison between MSR-GAN with different baselines in terms of relative-error versus SNR of the observations. In this experiment $d = 60$ and $m = 18$. All three methods have been initialized with the same signal and PMF and the reported results are the median across 10 different initializations and noise realizations for the observations.

the quantitative measure of the performance, defined as rel-error $= \frac{\min_s \|x - \mathcal{R}_s \widehat{x}\|^2}{\|x\|^2}$, where $\mathcal{R}_s$ shifts its input by $s \in \{0, ..., d-1\}$. To assess the quality of the estimated PMF, we use total variation (TV) distance, defined as TV $= \frac{1}{2} \min_s \|p - \mathcal{R}_s \widehat{p}\|_1$ [25]. We also define success rate by running MSR solutions with 10 different initializations. The ratio of the initializations that lead to a relative-error less than a threshold 0.02 is reported as the success-rate.

We compare MSR-GAN to two baselines: 1) Estimating shift-invariant features, i.e. moments up to the third order, from the measurements and recovering x and p by solving a non-convex optimization problem [7]. We use up to third order moments as the features. We call this baseline MSR via shift-invariant features (MSR-SIF). We use Riemannian trust-regions method [26] implemented in Manopt [27] to solve the optimization problem. 2) Expectation maximization (EM). In this baseline, we formulate MSR as a maximum marginalized likelihood estimation problem and solve it via EM [8, 7].

Effect of knowledge of PMF on the MSR-GAN results: Figure 3 shows the results of MSR-GAN on different signals with $d = 64$ and $m = 24$ in three different scenarios: 1) p in known (first column), 2) p is not known but fixed with a uniform distribution during training (second column), 3) p is not known and we recover it along side x (third and fourth columns). Note that for all three scenarios, we are using Alg. 1 with the same discriminator architecture and $\ell = 100$. However, for the first and the second scenarios, we do not update p (skip step 9-update p), rather keep it fixed with the true and the uniform distribution, respectively.

Note that when the PMF is known, the results of MSR-GAN closely match the ground truth signal. When the PMF is unknown, if we fix p to be a uniform distribution (see second column of Fig. 3), we observe that although the reconstructed signal is close to the GT, it has larger relative error compared to the scenarios where the PMF is given (see the first column of Fig. 3) or the PMF is updated jointly with the signal (see the third column of Fig. 3). Updating p jointly with x (Fig. 3-third column) leads to more accurate reconstruction of the signal. This shows the importance of recovering the distribution of the segments.

Effect of segment length and comparison with baselines: Figure 4 illustrates the effect of segment length m on the success rate of MSR-GAN compared to the other two baselines. For this experiment, we set the network hyper-parameter $\ell = 300$ and test the performance of our algorithm on randomly generated signals of length $d = 60$. As discussed in [7], solving MSR for smaller segment length regimes using shift invariant features is more challenging, as the number of equations provided by the moments for smaller segment lengths can be less than the number of unknowns. Similarly, the EM algorithm fails at shorter segments, i.e. $m \leq 25$, where the success rate is less than 50%. EM is more likely to get stuck at a local optimal solution when the segment length becomes smaller. However, as MSR-GAN solves the inverse problem by matching the distribution of real measurements and stochastic gradient descent, it achieves higher success rates for smaller segment lengths. In particular, even at $m = 15$, MSR-GAN achieves a success rate close to 100%.

Effect of noise and comparison with baselines: In Fig. 5, we investigate the effect of noise on the performance of MSR-GAN compared to the baselines. For this experiment $d = 60$, $m = 18$ and for the discriminator's architecture we set $\ell = 300$. Note that in different noise regimes MSR-GAN outperforms MSR-SIF and EM. Here we have a short segment length, thus as mentioned earlier solving MSR is more challenging and both baselines get stuck in local minima that is not close to the ground truth solution. Note that if we increase the segment length we observe an improved reconstruction error and success rate for MSR-SIF and EM (as also observed in Fig. 4). This suggests that MSR-GAN is a better solution compared to the baselines in short segment length regimes.

6. CONCLUSION

In this paper, we focused on the multi-segment reconstruction (MSR) problem, where we are given noisy randomly located segments of an unknown signal and the goal is to recover the signal and the distribution of the segments. We proposed a novel adversarial learning based approach to solve MSR. Our approach relies on distribution matching between the real measurements and the ones generated by the estimated signal and segment distribution. We formulated our problem in a Wasserstein GAN based framework. We showed how the generator loss term is a non-differentiable function of the segments distribution. To facilitate updates of the distribution through its gradients, we approximate the loss function at the generator side using Gumbel-Softmax reparametrization trick. This allowed us to update both the signal and the segment distribution using stochastic gradient descent. Our simulation results and comparisons to various baselines verified the ability of our approach in accurately solving MSR in various noise regimes and segment lengths.

7. REFERENCES

[1] H. Gupta, M. T. McCann, L. Donati, and M. Unser, "Cryogan: A new reconstruction paradigm for single-particle cryo-em via deep adversarial learning," *bioRxiv*, 2020.

[2] A. S. Motahari, G. Bresler, and D. N. C. Tse, "Information theory of DNA shotgun sequencing," *IEEE Transactions on Information Theory*, vol. 59, pp. 6273 – 6289, 2013.

[3] G. Paikin and A. Tal, "Solving multiple square jigsaw puzzles with missing pieces," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, 2015, pp. 161–174.

[4] M. Willemink and P. Noël, "The evolution of image reconstruction for ct—from filtered back projection to artificial intelligence," *European Radiology*, vol. 29, 10 2018.

[5] A. Barnett, L. Greengard, A. Pataki, and M. Spivak, "Rapid solution of the cryo-em reconstruction problem by frequency marching," *SIAM Journal on Imaging Sciences*, vol. 10, no. 3, pp. 1170–1195, 2017.

[6] A. Punjani, M. A. Brubaker, and D. J. Fleet, "Building proteins in a day: Efficient 3D molecular structure estimation with electron cryomicroscopy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 706–718, 2017.

[7] M. Zehni, M. N. Do, and Z. Zhao, "Multi-segment reconstruction using invariant features," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4629–4633.

[8] T. Bendory, N. Boumal, C. Ma, Z. Zhao, and A. Singer, "Bispectrum inversion with application to multireference alignment," *IEEE Transactions on signal processing*, vol. 66, no. 4, pp. 1037–1050, 2017.

[9] Y. Chen and E. Candès, "The projected power method: An efficient algorithm for joint alignment from pairwise differences," *Communications on Pure and Applied Mathematics*, vol. 71, 09 2016.

[10] S. Basu and Y. Bresler, "Feasibility of tomography with unknown view angles," *IEEE Transactions on Image Processing*, vol. 9, no. 6, pp. 1107–1122, Jun 2000.

[11] N. Boumal, T. Bendory, R. R. Lederman, and A. Singer, "Heterogeneous multireference alignment: a single pass approach," *ArXiv e-prints*, Oct. 2017.

[12] A. S. Bandeira, J. Niles-Weed, and P. Rigollet, "Optimal rates of estimation for multi-reference alignment," *Mathematical Statistics and Learning*, vol. 2, no. 1, pp. 25–75, 2020.

[13] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM REVIEW*, vol. 51, no. 3, pp. 455–500, 2009.

[14] R. Harshman, "Foundations of the parafac procedure: Models and conditions for an "explanatory" multi-modal factor analysis," *UCLA Working Papers in Phonetics*, vol. 16, 1970.

[15] H. Chen, M. Zehni, and Z. Zhao, "A spectral method for stable bispectrum inversion with application to multireference alignment," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 911–915, 2018.

[16] E. Abbe, J. M. Pereira, and A. Singer, "Sample complexity of the boolean multireference alignment problem," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 1316–1320.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., pp. 2672–2680. Curran Associates, Inc., 2014.

[18] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 214–223.

[19] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2017, NIPS'17, p. 5769–5779, Curran Associates Inc.

[20] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[21] C. J. Maddison, D. Tarlow, and T. Minka, "A$*$ sampling," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., pp. 3086–3094. Curran Associates, Inc., 2014.

[22] L. Bottou, "Large-scale machine learning with stochastic gradient descent," *Proc. of COMPSTAT*, 01 2010.

[23] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *International Conference on Learning Representations*, 2018.

[24] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.

[25] T. Chen and S. Kiefer, "On the total variation distance of labelled markov chains," in *Proceedings of the Joint Meeting of the Twenty-Third EACSL Annual Conference on Computer Science Logic (CSL) and the Twenty-Ninth Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, New York, NY, USA, 2014, CSL-LICS '14, Association for Computing Machinery.

[26] P.-A. Absil, C. G. Baker, and K. A. Gallivan, "Trust-region methods on Riemannian manifolds," *Found. Comput. Math.*, vol. 7, no. 3, pp. 303–330, July 2007.

[27] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre, "Manopt, a matlab toolbox for optimization on manifolds," *Journal of Machine Learning Research*, vol. 15, no. 42, pp. 1455–1459, 2014.