Asymptotics of MAP Inference in Deep Networks

Parthe Pandit, Mojtaba Sahraee, Allie K. Fletcher, Sundeep Rangan

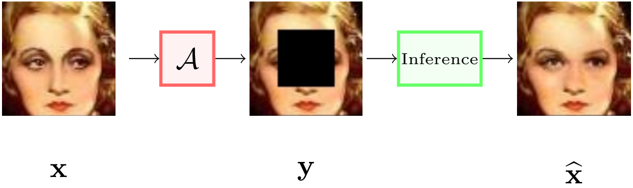
ECE, STAT — UCLA
ECE — NYU

ISIT 2019
9th July 2019

Outline

- Inference using Deep Generative Priors
  - Powerful reconstruction method
  - But difficult to analyze
- Framework: Approximate Message Passing
- Key Result: Exact prediction of performance in MAP inference
- Conclusions and Future work

Illustrative example: Inpainting



[YCL$^+$16, BJPD17]

Illustrative example: Inpainting

$\mathbf{x}$ $\qquad$ $\mathbf{y}$ $\qquad$ $\widehat{\mathbf{x}}$

P_X - *Prior* $\qquad$ $P_{Y|X}$ - *Likelihood*

Maximum a posteriori (MAP) inference

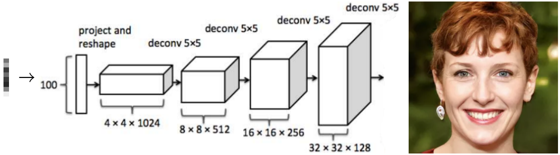
$$\widehat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmax}}\ P(\mathbf{x} \mid \mathbf{y}) = \underset{\mathbf{x}}{\operatorname{argmax}}\ P(\mathbf{x})\, P_{\mathcal{A}}(\mathbf{y} \mid \mathbf{x})$$

[YCL+16, BJPD17]

But.. what is a good prior for human faces?

$\mathbf{x} \sim$ 

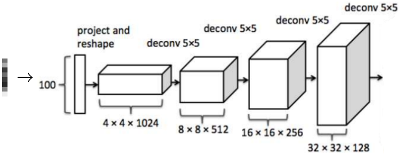
Deep Generative Priors



$\mathbf{z}$ $\quad\quad$ $\mathcal{G}$ $\quad\quad$ $\mathbf{x}$

- Priors *learned* from data
- $\mathbf{x} = \mathcal{G}(\mathbf{z})$ is a function of a *simple* random variable

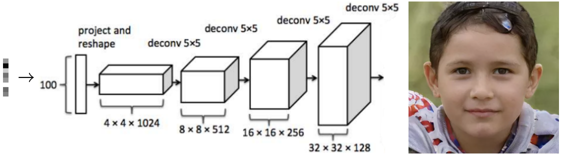
Deep Generative Priors



$\mathbf{z}$ $\mathcal{G}$ $\mathbf{x}$

- Priors *learned* from data
- $\mathbf{x} = \mathcal{G}(\mathbf{z})$ is a function of a *simple* random variable
- Example: $\mathcal{G}$ is a neural network and $\mathbf{z} \sim \mathcal{N}(0, I_d)$

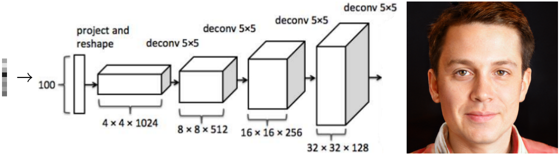
Deep Generative Priors



$\mathbf{z}$ $\mathcal{G}$ $\mathbf{x}$

- Priors *learned* from data
- $\mathbf{x} = \mathcal{G}(\mathbf{z})$ is a function of a *simple* random variable
- Example: $\mathcal{G}$ is a neural network and $\mathbf{z} \sim \mathcal{N}(0, I_d)$

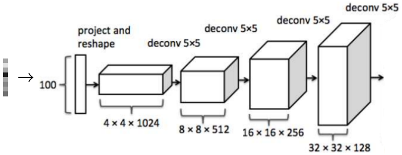
Deep Generative Priors



$\mathbf{z}$ $\mathcal{G}$ $\mathbf{x}$

- Priors *learned* from data
- $\mathbf{x} = \mathcal{G}(\mathbf{z})$ is a function of a *simple* random variable
- Example: $\mathcal{G}$ is a neural network and $\mathbf{z} \sim \mathcal{N}(0, I_d)$

Deep Generative Priors



$\mathbf{z}$ $\mathcal{G}$ $\mathbf{x}$

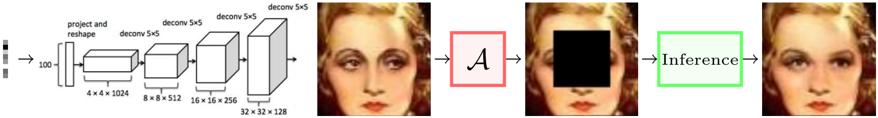
- Priors *learned* from data
- $\mathbf{x} = \mathcal{G}(\mathbf{z})$ is a function of a *simple* random variable
- Example: $\mathcal{G}$ is a neural network and $\mathbf{z} \sim \mathcal{N}(0, I_d)$
- Train $\mathcal{G}$ on a dataset using some *loss function*
  - VAE Variational Autoencoder [KW13]
  - GAN Generative Adversarial Network [GBC16]
  - Variants of these two: WGAN[ACB17], DCGAN [RMC15],...

https://thispersondoesnotexist.com

$\mathbf{x} \quad \sim$

Generated using StyleGAN trained on ≈ 200k images [KLA19]

Inference using Deep Generative Priors



$\mathbf{z}$ $\mathcal{G}$ $\mathbf{x}$ $\mathbf{y}$ $\widehat{\mathbf{x}}$

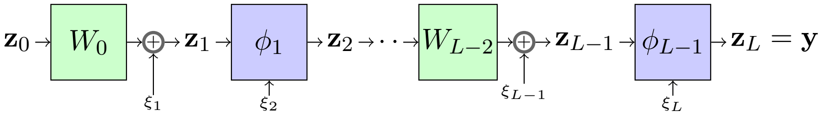
$P_{Z,X}$ - *Prior* $P_{Y|X}$ - *Likelihood*

MAP inference

$$\underset{\mathbf{z},\mathbf{x}}{\operatorname{argmax}}\; P(\mathbf{z},\mathbf{x} \mid \mathbf{y}) = \underset{\mathbf{z},\mathbf{x}}{\operatorname{argmax}}\; P(\mathbf{z}) P_{\mathcal{G}}(\mathbf{x} \mid \mathbf{z}) P_{\mathcal{A}}(\mathbf{y} \mid \mathbf{x})$$

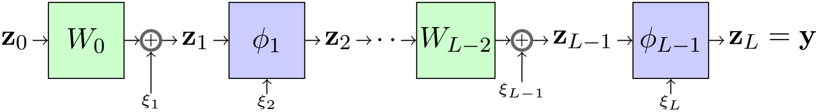
[RMC15, YCL^{+}16, BJPD17]

MAP Inference in Deep Networks



- Problem: Estimate $\mathbf{z} := \{\mathbf{z}_\ell\}_{\ell=0}^{L-1}$ from $\mathbf{y}$ and$\{W_{2\ell}\}$

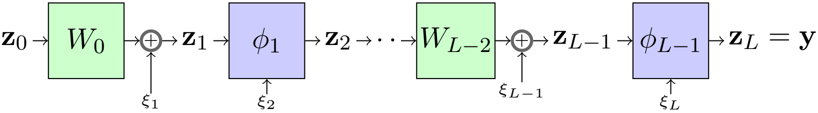
MAP Inference in Deep Networks



- Problem: Estimate $\mathbf{z} := \{\mathbf{z}_\ell\}_{\ell=0}^{L-1}$ from $\mathbf{y}$ and$\{W_{2\ell}\}$
- MAP inference:

$$\operatorname*{argmin}_{\mathbf{z}} F(\mathbf{z}) = -\ln P(\mathbf{z}_0) - \sum_{\ell} \ln P(\mathbf{z}_\ell \mid \mathbf{z}_{\ell-1})$$

MAP Inference in Deep Networks

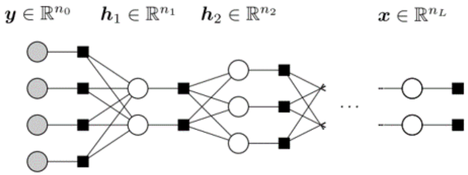


- Problem: Estimate $\mathbf{z} := \{\mathbf{z}_\ell\}_{\ell=0}^{L-1}$ from $\mathbf{y}$ and $\{W_{2\ell}\}$
- MAP inference:

$$\operatorname*{argmin}_{\mathbf{z}} F(\mathbf{z}) = -\ln P(\mathbf{z}_0) - \sum_{\ell} \ln P(\mathbf{z}_\ell \mid \mathbf{z}_{\ell-1})$$

- Scaling laws have been studied in [HV17, SH18]
- This work: Exact prediction of MSE: $\frac{1}{n_\ell} \left\|\widehat{\mathbf{z}}_\ell - \mathbf{z}_\ell^*\right\|^2$

Previous Work in AMP for Inference in Deep networks



[MKMZ17]

- AMP algorithm [DMM09] for LASSO
  - Predictable reconstruction error for i.i.d. Gaussian design
  - $\mathcal{E}_{t+1} = \mathbb{E}\left\{g\left(X_o + Z\sqrt{\frac{n_{\text{in}}}{n_{\text{out}}}\mathcal{E}_t}\right) - X_o\right\}^2$
  - State Evolution has been rigourously proven [BM11]
  - Vector AMP [RSF17] extends to rotationally invariant design
- Multi-layer AMP for MMSE [MKMZ17, GML+18, Ree17]
- Multi-layer Vector AMP for MMSE [FRS18]
  - Propose SE, derives the optimal MSE

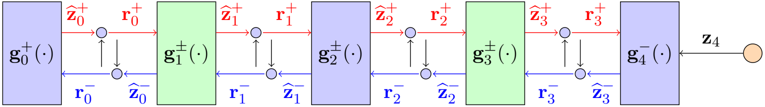
Our contribution

- MAP is more computationally tractable than MMSE
  - Successfully used by practical solvers [BJPD17, YCL+16, CLPK17]
  - Implemented in most deep learning packages (e.g. Tensorflow)

- This work: MAP inference using MLVAMP methodology
- What we show:
  1. Fixed points of ML-VAMP are KKT points of MAP
  2. Exact asymptotic MSE of ML-VAMP trajectories

- Key takeaway: A MAP algorithm with rigourous analysis

Multi-layer Vector Approximate Message Passing

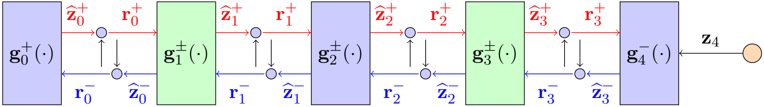


- MLVAMP estimates $\mathbf{z}_\ell$ as a sequence of proximal operations

$$(\widehat{\mathbf{z}}_\ell^+, \widehat{\mathbf{z}}_{\ell-1}^-) = \underset{\mathbf{z}_\ell^+, \mathbf{z}_{\ell-1}^-}{\operatorname{argmin}} \; -\ln P(\mathbf{z}_\ell^+ \mid \mathbf{z}_{\ell-1}^-) + \frac{\gamma_\ell^-}{2} \left\| \mathbf{z}_\ell^+ - \mathbf{r}_\ell^- \right\|^2 + \frac{\gamma_{\ell-1}^+}{2} \left\| \mathbf{z}_{\ell-1}^- - \mathbf{r}_{\ell-1}^+ \right\|^2$$

- $\{\mathbf{r}_\ell^\pm, \gamma_\ell^\pm\}$ updated based on factor graph derivation
- Multilayer extension of expectation consistent inference [HWJ17, OW05]

Multi-layer Vector Approximate Message Passing



- MLVAMP estimates $\mathbf{z}_\ell$ as a sequence of proximal operations

$$(\widehat{\mathbf{z}}_\ell^+, \widehat{\mathbf{z}}_{\ell-1}^-) = \underset{\mathbf{z}_\ell^+, \mathbf{z}_{\ell-1}^-}{\operatorname{argmin}} \; -\ln P(\mathbf{z}_\ell^+ \mid \mathbf{z}_{\ell-1}^-) + \frac{\gamma_\ell^-}{2}\left\|\mathbf{z}_\ell^+ - \mathbf{r}_\ell^-\right\|^2 + \frac{\gamma_{\ell-1}^+}{2}\left\|\mathbf{z}_{\ell-1}^- - \mathbf{r}_{\ell-1}^+\right\|^2$$

Theorem (1)

(i) Fixed points are critical points of augmented Lagrangian

$$\mathcal{L}(\mathbf{z}^+, \mathbf{z}^-, \lambda) := F(\mathbf{z}^+, \mathbf{z}^-) + \lambda^\top(\mathbf{z}^+ - \mathbf{z}^-) + \frac{\eta}{2}\left\|\mathbf{z}^+ - \mathbf{z}^-\right\|^2$$

(ii) Fixed $\gamma_\ell^\pm$: MLVAMP matches Peaceman Rachford splitting ADMM

$$\widehat{\mathbf{z}}^+ \leftarrow \operatorname{argmin}_{\mathbf{z}^+} \; \mathcal{L}(\mathbf{z}^+, \widehat{\mathbf{z}}^-, \lambda) \qquad \lambda_\ell^+ \leftarrow \lambda_\ell^- + \gamma_\ell^+/\eta_\ell(\widehat{\mathbf{z}}_\ell^+ - \widehat{\mathbf{z}}_\ell^-)$$

$$\widehat{\mathbf{z}}^- \leftarrow \operatorname{argmin}_{\mathbf{z}^-} \; \mathcal{L}(\widehat{\mathbf{z}}^+, \mathbf{z}^-, \lambda) \qquad \lambda_\ell^- \leftarrow \lambda_\ell^+ + \gamma_\ell^-/\eta_\ell(\widehat{\mathbf{z}}_\ell^+ - \widehat{\mathbf{z}}_\ell^-)$$

Exact prediction of MAP reconstruction error

- All hidden dimensions $n_\ell \to \infty$ such that $\frac{n_\ell}{n_0} \to \beta_\ell = \Theta(1)$
- Weight matrices are rotationally invariant and independent
  - Richer ensemble than i.i.d. Gaussian matrices
  - Singular value spectrum can be arbitrary but i.i.d.
  - Generalizes beyond the Marchenko-Pastur law
- Input $\mathbf{z}_0$, noise ξ_ℓ are iid distributed

Theorem (2)

For every iteration $t \geq 1$, and for all layers $\ell = 0, 1, \ldots, L-1$

$$\tau^+_{\ell,t} := \lim_{n_\ell \to \infty} \frac{1}{n_\ell} \left\| \widehat{\mathbf{z}}^+_{\ell t} - \mathbf{z}^*_\ell \right\|_2^2 = \mathcal{E}^+_\ell(\tau^-_{\ell,t}, \tau^+_{\ell-1,t})$$

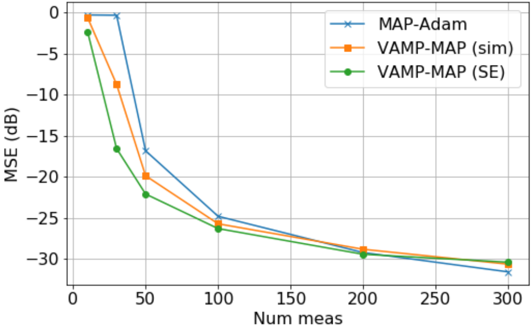
$$\tau^-_{\ell,t} := \lim_{n_\ell \to \infty} \frac{1}{n_\ell} \left\| \widehat{\mathbf{z}}^-_{\ell,t} - \mathbf{z}^*_\ell \right\|_2^2 = \mathcal{E}^-_\ell(\tau^-_{\ell+1,t}, \tau^+_{\ell,t-1})$$

are recursively computable from scalar laws $\{\mathbb{P}_{S_{2\ell}}, \mathbb{P}_{\xi_\ell}\}$, and $\mathbb{P}_{Z_0}$.

MLVAMP Example 1: Synthetic Network ($L = 6$)

Input layer $\mathbb{R}^{20}$. Hidden layers $\mathbb{R}^{100}$ and $\mathbb{R}^{500}$

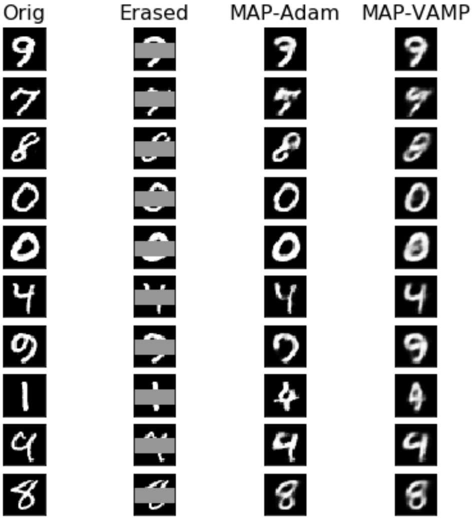
$$10\log_{10}\frac{\left\|\widehat{\mathbf{z}}_0-\mathbf{z}_0^*\right\|^2}{\left\|\mathbf{z}_0^*\right\|^2}$$



size of output layer

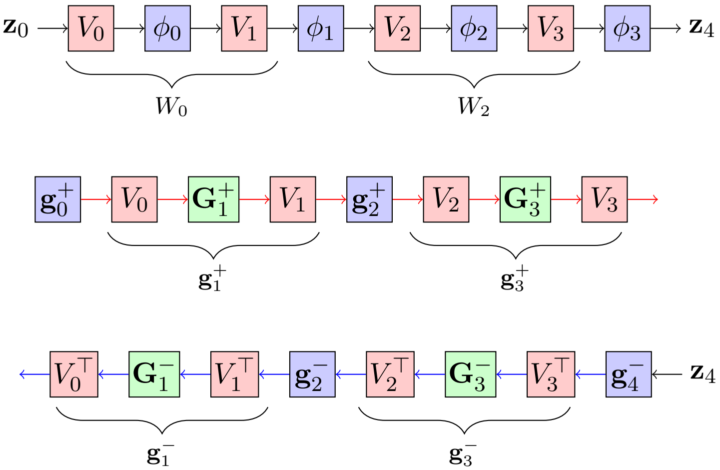
- Performs comparable to ADAM
- Performance of ML-VAMP can be predicted exactly

MLVAMP Example 2: MNIST inpainting



- VAE generative prior with
  - Input layer $\mathbb{R}^{20}$
  - Hidden layer $\mathbb{R}^{100}$
  - Output layer $\mathbb{R}^{28\times28}$
  - 50k training images
- Implemented using Tensorflow
- Performs comparable to ADAM

Sketch of the Proof of State Evolution

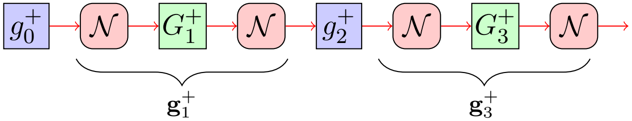


Sketch of the Proof of State Evolution

- Key Insight: Random unitary V act like Gaussian sampler.

$$\mathbf{x} \to \boxed{V} \to \mathbf{y} \quad \equiv \quad X \to \boxed{\mathcal{N}} \to Y$$

- Asymptotics for first forward pass



- For later iterations inputs to V_ℓ depend on V_ℓ
- Bolthausen Conditioning: [BM11]

Concluding remarks

- Reconstruction with deep generative priors
  - Incredible practical results
  - But difficult to analyze
- MAP inference using MLVAMP methodology
  - Finds stationary points of MAP optimization problem
  - Computationally tractable algorithm
  - Enables exact prediction of performance via State Evolution

Future directions

- Convolutional layers, resnet
- Stability: Algorithm requires damping for convergence
- Networks with learning:
  - Current problem is on pre-trained networks
  - Can use parameter estimation similar to other VAMP works
  - Or, treat layers with joint unknowns of weights and inputs

Martin Arjovsky, Soumith Chintala, and Léon Bottou.
Wasserstein GAN.
arXiv:1701.07875, 2017.

Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis.
Compressed sensing using generative models.
ICML, 2017.

M. Bayati and A. Montanari.
The dynamics of message passing on dense graphs, with applications to compressed sensing.
IEEE Trans. Info. Theory, 57(2):764–785, Feb. 2011.

JH Rick Chang, Chun-Liang Li, Barnabas Poczos, and BVK Vijaya Kumar.
One network to solve them all—solving linear inverse problems using deep projection models.
In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5889–5898. IEEE, 2017.

D. L. Donoho, A. Maleki, and A. Montanari.
Message-passing algorithms for compressed sensing.
PNAS, 106(45):18914–18919, Nov. 2009.

Alyson K Fletcher, Sundeep Rangan, and P. Schniter.
Inference in deep networks in high dimensions.
Proc. IEEE ISIT, 2018.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville.
Deep learning.
MIT Press, 2016.

Marylou Gabrié, Andre Manoel, Clément Luneau, Jean Barbier, Nicolas Macris, Florent Krzakala, and Lenka Zdeborová.
Entropy and mutual information in models of deep neural networks.
NIPS, 2018.

Paul Hand and Vladislav Voroninski.

Global guarantees for enforcing deep generative priors by empirical risk.
arXiv:1705.07576, 2017.

Hengtao He, Chao-Kai Wen, and Shi Jin.
Generalized expectation consistent signal recovery for nonlinear measurements.
In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2333–2337. IEEE, 2017.

Tero Karras, Samuli Laine, and Timo Aila.
A style-based generator architecture for generative adversarial networks.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

Diederik P Kingma and Max Welling.
Auto-encoding variational bayes.
arXiv:1312.6114, 2013.

Andre Manoel, Florent Krzakala, Marc Mézard, and Lenka Zdeborová.
Multi-layer generalized linear estimation.
arXiv:1701.06981, 2017.

Manfred Opper and Ole Winther.
Expectation consistent approximate inference.
Journal of Machine Learning Research, 6(Dec):2177–2204, 2005.

Galen Reeves.
Additivity of information in multilayer networks via additive gaussian noise transforms.
arXiv:1710.04580, 2017.

Alec Radford, Luke Metz, and Soumith Chintala.
Unsupervised representation learning with deep convolutional generative adversarial networks.
arXiv:1511.06434, 2015.

Sundeep Rangan, Philip Schniter, and Alyson K. Fletcher.

Vector approximate message passing.
In *Proc. IEEE ISIT*, pages 1588–1592, 2017.

Viraj Shah and Chinmay Hegde.
Solving linear inverse problems using gan priors: An algorithm with provable guarantees.
arXiv:1802.08406, 2018.

Raymond Yeh, Chen Chen, Teck Yian Lim, Mark Hasegawa-Johnson, and Minh N Do.
Semantic image inpainting with perceptual and contextual losses.
arXiv:1607.07539, 2016.