On the Statistical Rate of Nonlinear Recovery in Generative Models with Heavy-tailed Data

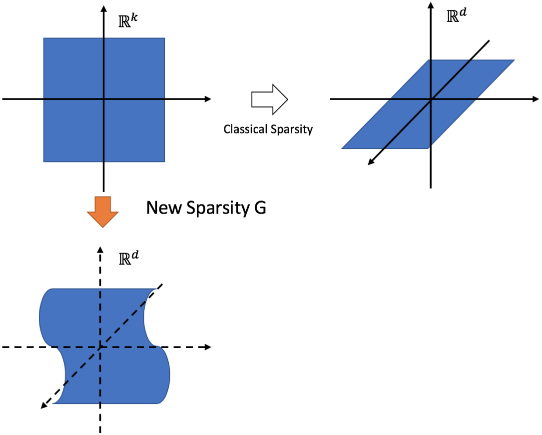
Xiaohan Wei, Zhuoran Yang, and Zhaoran Wang

University of Southern California, Princeton University and Northwestern University

June 12th, 2019

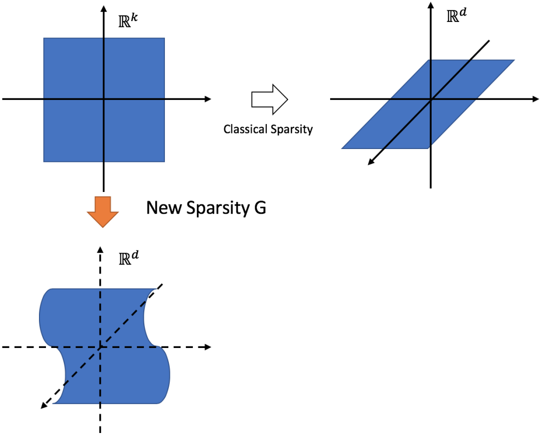
Generative Model vs Sparsity in Signal Recovery

- Classical sparsity: structure of the signals depend on basis.



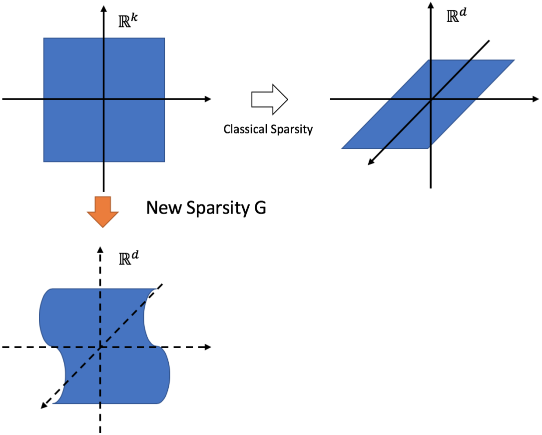
Generative Model vs Sparsity in Signal Recovery

- Classical sparsity: structure of the signals depend on basis.
- Generative model: explicit parametrization of low-dimensional signal manifold.

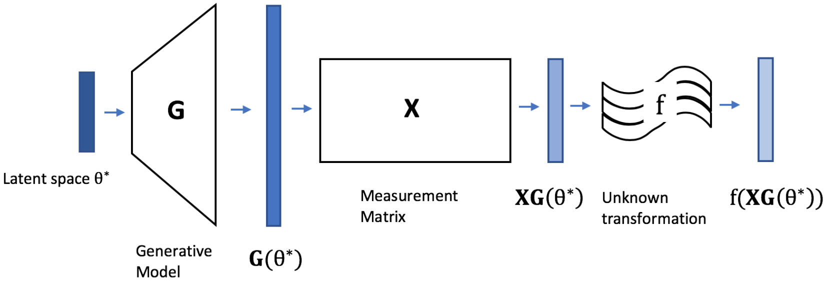


Generative Model vs Sparsity in Signal Recovery

- Classical sparsity: structure of the signals depend on basis.
- Generative model: explicit parametrization of low-dimensional signal manifold.
- Previous works: [Bora et al. 2017] [Hand et al. 2018] [Mardani et al. 2017].

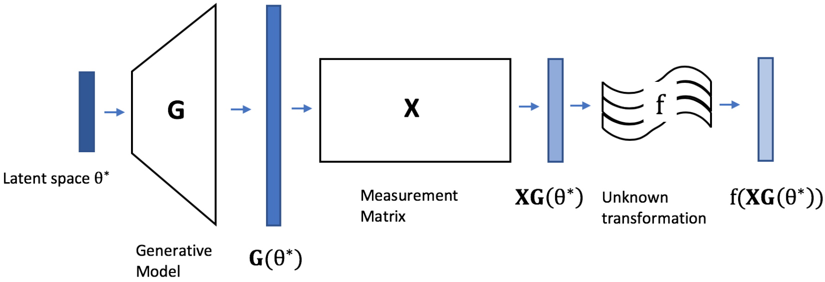


Nonlinear Recovery via Generative Models



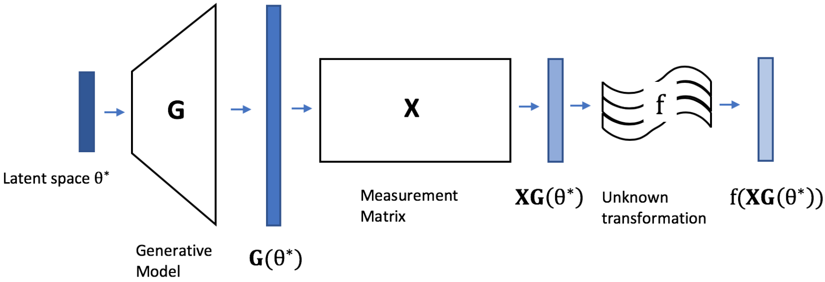
- Given: Generative model $\mathbf{G} : \mathbb{R}^k \to \mathbb{R}^d$ and measurement matrix $\mathbf{X} \in \mathbb{R}^{m \times d}$.

Nonlinear Recovery via Generative Models



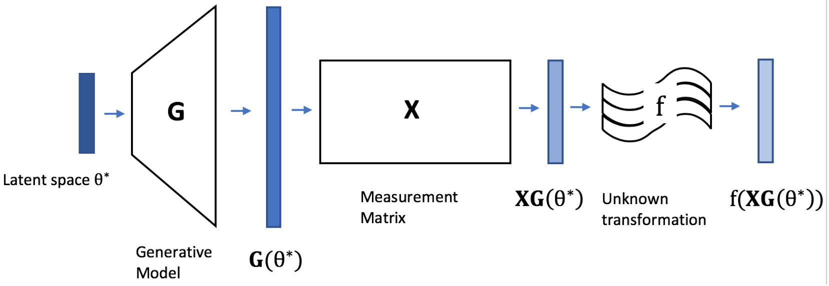
- Given: Generative model $\mathbf{G} : \mathbb{R}^k \to \mathbb{R}^d$ and measurement matrix $\mathbf{X} \in \mathbb{R}^{m \times d}$.
- Goal: Recovery $\mathbf{G}(\theta^*)$ up to scaling from nonlinear observations $\mathbf{y} = f(\mathbf{X}\mathbf{G}(\theta^*))$.

Nonlinear Recovery via Generative Models



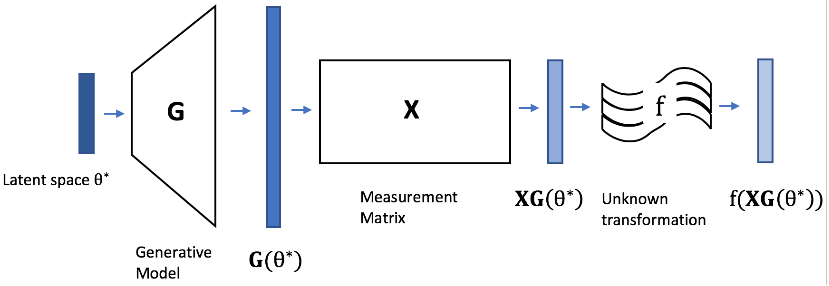
- Given: Generative model $\mathbf{G} : \mathbb{R}^k \to \mathbb{R}^d$ and measurement matrix $\mathbf{X} \in \mathbb{R}^{m\times d}$.
- Goal: Recovery $\mathbf{G}(\theta^*)$ up to scaling from nonlinear observations $\mathbf{y} = f(\mathbf{XG}(\theta^*))$.
- Challenges:
  1. High-dimensional recovery: $k \ll d$, $m \ll d$.

Nonlinear Recovery via Generative Models



- Given: Generative model $\mathbf{G} : \mathbb{R}^k \to \mathbb{R}^d$ and measurement matrix $\mathbf{X} \in \mathbb{R}^{m \times d}$.
- Goal: Recovery $\mathbf{G}(\theta^*)$ up to scaling from nonlinear observations $\mathbf{y} = f(\mathbf{XG}(\theta^*))$.
- Challenges:
  1. High-dimensional recovery: $k \ll d$, $m \ll d$.
  2. Non-Gaussian $\mathbf{X}$ and *unknown* non-linearity f.

Nonlinear Recovery via Generative Models



- Given: Generative model $\mathbf{G} : \mathbb{R}^k \to \mathbb{R}^d$ and measurement matrix $\mathbf{X} \in \mathbb{R}^{m \times d}$.
- Goal: Recovery $\mathbf{G}(\theta^*)$ up to scaling from nonlinear observations $\mathbf{y} = f(\mathbf{XG}(\theta^*))$.
- Challenges:
  1. High-dimensional recovery: $k \ll d$, $m \ll d$.
  2. Non-Gaussian $\mathbf{X}$ and *unknown* non-linearity f.
  3. Observations $\mathbf{y}$ can be *heavy-tailed*.

Our Method: Stein + Adaptive Thresholding

- Suppose the rows of $\mathbf{X} := [\mathbf{X}_1, \cdots, \mathbf{X}_m]^T \in \mathbb{R}^{m \times d}$ have density $p : \mathbb{R}^d \to \mathbb{R}$.
- Define the (row-wise) score transformation:

$$\mathcal{S}_p(\mathbf{X}) := [\mathcal{S}_p(\mathbf{X}_1), \cdots, \mathcal{S}_p(\mathbf{X}_m)]^T = [\nabla \log p(\mathbf{X}_1), \cdots, \nabla \log p(\mathbf{X}_m)]^T.$$

Our Method: Stein + Adaptive Thresholding

- Suppose the rows of $\mathbf{X} := [\mathbf{X}_1, \cdots, \mathbf{X}_m]^T \in \mathbb{R}^{m \times d}$ have density $p : \mathbb{R}^d \to \mathbb{R}$.
- Define the (row-wise) score transformation:

$$\mathcal{S}_p(\mathbf{X}) := [\mathcal{S}_p(\mathbf{X}_1), \cdots, \mathcal{S}_p(\mathbf{X}_m)]^T = [\nabla \log p(\mathbf{X}_1), \cdots, \nabla \log p(\mathbf{X}_m)]^T.$$

- (First-order) Stein's identity: when $\mathbb{E} f'(\langle \mathbf{X}_i, \mathbf{G}(\theta^*) \rangle) > 0$,

$$\mathbb{E}\left[\mathcal{S}_p(\mathbf{X})^T \mathbf{y}\right] \propto \mathbf{G}(\theta^*).$$

- (Second-order) Stein's identity: when $\mathbb{E} f''(\langle \mathbf{X}_i, \mathbf{G}(\theta^*) \rangle) > 0$, δ is a constant,

$$\mathbb{E}\left[\mathcal{S}_p(\mathbf{X})^T \text{diag}(\mathbf{y}) \mathcal{S}_p(\mathbf{X})\right] \propto \mathbf{G}(\theta^*)\mathbf{G}(\theta^*)^T + \delta \cdot \mathbf{I}_{d \times d}.$$

Our Method: Stein + Adaptive Thresholding

- Suppose the rows of $\mathbf{X} := [\mathbf{X}_1, \cdots, \mathbf{X}_m]^T \in \mathbb{R}^{m \times d}$ have density $p : \mathbb{R}^d \to \mathbb{R}$.
- Define the (row-wise) score transformation:

$$\mathcal{S}_p(\mathbf{X}) := [\mathcal{S}_p(\mathbf{X}_1), \cdots, \mathcal{S}_p(\mathbf{X}_m)]^T = [\nabla \log p(\mathbf{X}_1), \cdots, \nabla \log p(\mathbf{X}_m)]^T.$$

- (First-order) Stein's identity: when $\mathbb{E} f'(\langle \mathbf{X}_i, \mathbf{G}(\theta^*) \rangle) > 0$,

$$\mathbb{E}\left[\mathcal{S}_p(\mathbf{X})^T \mathbf{y}\right] \propto \mathbf{G}(\theta^*).$$

- (Second-order) Stein's identity: when $\mathbb{E} f''(\langle \mathbf{X}_i, \mathbf{G}(\theta^*) \rangle) > 0$, δ is a constant,

$$\mathbb{E}\left[\mathcal{S}_p(\mathbf{X})^T \mathrm{diag}(\mathbf{y}) \mathcal{S}_p(\mathbf{X})\right] \propto \mathbf{G}(\theta^*) \mathbf{G}(\theta^*)^T + \delta \cdot \mathbf{I}_{d \times d}.$$

- Adaptive thresholding: suppose $\|y_i\|_{L_q} < \infty$, $q > 4$, and $\tau_m \propto m^{2/q}$,

$$\widetilde{y}_i = \mathrm{sign}(y_i) \cdot (|y_i| \wedge \tau_m), \quad i \in \{1, 2, \cdots, m\}$$

Our Method: Stein + Adaptive Thresholding

- Least-squares estimator:

$$\widehat{\theta} \in \operatorname{argmin}_{\theta \in \mathbb{R}^k} \left\| \mathbf{G}(\theta) - \frac{1}{m} \mathcal{S}_p(\mathbf{X})^T \widetilde{\mathbf{y}} \right\|_2^2 .$$

Our Method: Stein + Adaptive Thresholding

- Least-squares estimator:

$$\widehat{\theta} \in \text{argmin}_{\theta \in \mathbb{R}^k} \left\| \mathbf{G}(\theta) - \frac{1}{m} \mathcal{S}_p(\mathbf{X})^T \widetilde{\mathbf{y}} \right\|_2^2 .$$

- Main performance theorem:

Theorem (Wei, Yang and Wang, 2019)

For any accuracy level $\varepsilon \in (0,1]$*, suppose*
(1) $\mathbb{E} f'(\langle \mathbf{X}_i, \mathbf{G}(\theta^*) \rangle) > 0$*,*
(2) the generative model $\mathbf{G}$ *is a ReLU network with zero bias,*
(3) the number of measurements

$$m \propto k \varepsilon^{-2} \log d .$$

Then, with high probability,

$$\left\| \frac{\mathbf{G}(\widehat{\theta})}{\|\mathbf{G}(\widehat{\theta})\|_2} - \frac{\mathbf{G}(\theta^*)}{\|\mathbf{G}(\theta^*)\|_2} \right\|_2 \leq \varepsilon .$$

- Similar results hold for more general Lipschitz generators $\mathbf{G}$.

Our Method: Stein + Adaptive Thresholding

- PCA type estimator:

$$\widehat{\theta} \in \text{argmax}_{\|\mathbf{G}(\theta)\|_2=1} \; \mathbf{G}(\theta)^T \mathcal{S}_p(\mathbf{X})^T \text{diag}(\widetilde{\mathbf{y}}) \mathcal{S}_p(\mathbf{X}) \mathbf{G}(\theta)$$

Our Method: Stein + Adaptive Thresholding

- PCA type estimator:

$$\widehat{\theta} \in \operatorname{argmax}_{\|\mathbf{G}(\theta)\|_2=1} \ \mathbf{G}(\theta)^T \mathcal{S}_p(\mathbf{X})^T \mathrm{diag}(\widetilde{\mathbf{y}}) \mathcal{S}_p(\mathbf{X}) \mathbf{G}(\theta)$$

- Main performance theorem:

Theorem (Wei, Yang and Wang, 2019)

For any accuracy level $\varepsilon \in (0, 1]$, *suppose*
(1) $\mathbb{E} f''(\langle \mathbf{X}_i, \mathbf{G}(\theta^*) \rangle) > 0$,
(2) the generative model **G** *is a ReLU network with zero bias,*
(3) the number of measurements

$$m \propto k\varepsilon^{-2} \log d.$$

Then, with high probability,

$$\left\| \mathbf{G}(\widehat{\theta}) - \frac{\mathbf{G}(\theta^*)}{\|\mathbf{G}(\theta^*)\|_2} \right\|_2 \leq \varepsilon.$$

- Similar results hold for more general Lipschitz generators **G**.

Thank you!

Poster 198, Pacific Ballroom, 6:30-9:00 pm