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 \( \mathbf{X} \in \mathbb{R}^{m \times d} \).

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.

\[ \mathbf{y} = f(\mathbf{XG}(\theta^*)) \]
Given: Generative model $G : \mathbb{R}^k \to \mathbb{R}^d$ and measurement matrix $X \in \mathbb{R}^{m \times d}$.

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

Goal: Recovery $G(\theta^*)$ up to scaling from nonlinear observations $y = f(XG(\theta^*))$.

Challenges:
- High-dimensional recovery: $k \ll d$, $m \ll d$. 
Given: Generative model $G : \mathbb{R}^k \rightarrow \mathbb{R}^d$ and measurement matrix $X \in \mathbb{R}^{m \times d}$.

Goal: Recovery $G(\theta^*)$ up to scaling from nonlinear observations $y = f(XG(\theta^*))$.

Challenges:

1. High-dimensional recovery: $k \ll d$, $m \ll d$.
2. Non-Gaussian $X$ and unknown non-linearity $f$. 
Nonlinear Recovery via Generative Models

Given: Generative model \( G : \mathbb{R}^k \rightarrow \mathbb{R}^d \) and measurement matrix \( X \in \mathbb{R}^{m \times d} \).

Goal: Recovery \( G(\theta^*) \) up to scaling from nonlinear observations \( y = f(XG(\theta^*)) \).

Challenges:
1. High-dimensional recovery: \( k \ll d \), \( m \ll d \).
2. Non-Gaussian \( X \) and unknown non-linearity \( f \).
3. Observations \( y \) can be heavy-tailed.
Our Method: Stein + Adaptive Thresholding

- Suppose the rows of $X := [X_1, \cdots, X_m]^T \in \mathbb{R}^{m \times d}$ have density $p : \mathbb{R}^d \rightarrow \mathbb{R}$.
- Define the (row-wise) score transformation:

$$S_p(X) := [S_p(X_1), \cdots, S_p(X_m)]^T = [\nabla \log p(X_1), \cdots, \nabla \log p(X_m)]^T.$$
Our Method: Stein + Adaptive Thresholding

- Suppose the rows of $X := [X_1, \cdots, X_m]^T \in \mathbb{R}^{m \times d}$ have density $p : \mathbb{R}^d \to \mathbb{R}$.
- Define the (row-wise) score transformation:
  $$S_p(X) := [S_p(X_1), \cdots, S_p(X_m)]^T = [\nabla \log p(X_1), \cdots, \nabla \log p(X_m)]^T.$$
- (First-order) Stein's identity: when $\mathbb{E}f'(\langle X_i, G(\theta^*) \rangle) > 0$,
  $$\mathbb{E} \left[ S_p(X)^T y \right] \propto G(\theta^*).$$
- (Second-order) Stein's identity: when $\mathbb{E}f''(\langle X_i, G(\theta^*) \rangle) > 0$, $\delta$ is a constant,
  $$\mathbb{E} \left[ S_p(X)^T \text{diag}(y) S_p(X) \right] \propto G(\theta^*)G(\theta^*)^T + \delta \cdot I_{d \times d}.$$
Our Method: Stein + Adaptive Thresholding

- Suppose the rows of $X := [X_1, \cdots, X_m]^T \in \mathbb{R}^{m \times d}$ have density $p : \mathbb{R}^d \to \mathbb{R}$.
- Define the (row-wise) score transformation:
  $$S_p(X) := [S_p(X_1), \cdots, S_p(X_m)]^T = [\nabla \log p(X_1), \cdots, \nabla \log p(X_m)]^T.$$
- (First-order) Stein’s identity: when $\mathbb{E} f'(\langle X_i, G(\theta^*) \rangle) > 0$,
  $$\mathbb{E} \left[ S_p(X)^T y \right] \propto G(\theta^*).$$
- (Second-order) Stein’s identity: when $\mathbb{E} f''(\langle X_i, G(\theta^*) \rangle) > 0$, $\delta$ is a constant,
  $$\mathbb{E} \left[ S_p(X)^T \text{diag}(y) S_p(X) \right] \propto G(\theta^*)G(\theta^*)^T + \delta \cdot I_{d \times d}.$$
- Adaptive thresholding: suppose $\|y_i\|_{L_q} < \infty$, $q > 4$, and $\tau_m \propto m^2/q$,
  $$\tilde{y}_i = \text{sign}(y_i) \cdot (|y_i| \wedge \tau_m), \quad i \in \{1, 2, \cdots, m\}.$$
Our Method: Stein + Adaptive Thresholding

- Least-squares estimator:

\[
\hat{\theta} \in \text{argmin}_{\theta \in \mathbb{R}^k} \left\| G(\theta) - \frac{1}{m} S_p(X)^T \tilde{y} \right\|_2^2.
\]
Our Method: Stein + Adaptive Thresholding

- Least-squares estimator:

\[ \hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^k} \left\| G(\theta) - \frac{1}{m} S_p(X)^T \tilde{y} \right\|_2^2. \]

- Main performance theorem:

**Theorem (Wei, Yang and Wang, 2019)**

For any accuracy level \( \varepsilon \in (0, 1) \), suppose

1. \( \mathbb{E} f'(\langle X_i, G(\theta^*) \rangle) > 0 \),
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,

\[ \left\| \frac{G(\hat{\theta})}{\|G(\hat{\theta})\|_2} - \frac{G(\theta^*)}{\|G(\theta^*)\|_2} \right\|_2 \leq \varepsilon. \]

- Similar results hold for more general Lipschitz generators \( G \).
Our Method: Stein + Adaptive Thresholding

- PCA type estimator:

\[
\hat{\theta} \in \arg\max_{\|G(\theta)\|_2=1} G(\theta)^T S_p(X)^T \text{diag}(\tilde{y}) S_p(X) G(\theta)
\]
Our Method: Stein + Adaptive Thresholding

- PCA type estimator:

\[
\hat{\theta} \in \arg\max_{\|G(\theta)\|_2 = 1} G(\theta)^T S_p(X)^T \text{diag}(\tilde{y}) S_p(X) G(\theta)
\]

- Main performance theorem:

**Theorem (Wei, Yang and Wang, 2019)**

For any accuracy level \( \varepsilon \in (0, 1] \), suppose

1. \( \mathbb{E} f''(\langle X_i, G(\theta^*) \rangle) > 0 \),
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,

\[
\left\| G(\hat{\theta}) - \frac{G(\theta^*)}{\|G(\theta^*)\|_2} \right\|_2 \leq \varepsilon.
\]

- Similar results hold for more general Lipschitz generators \( G \).
Thank you!

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