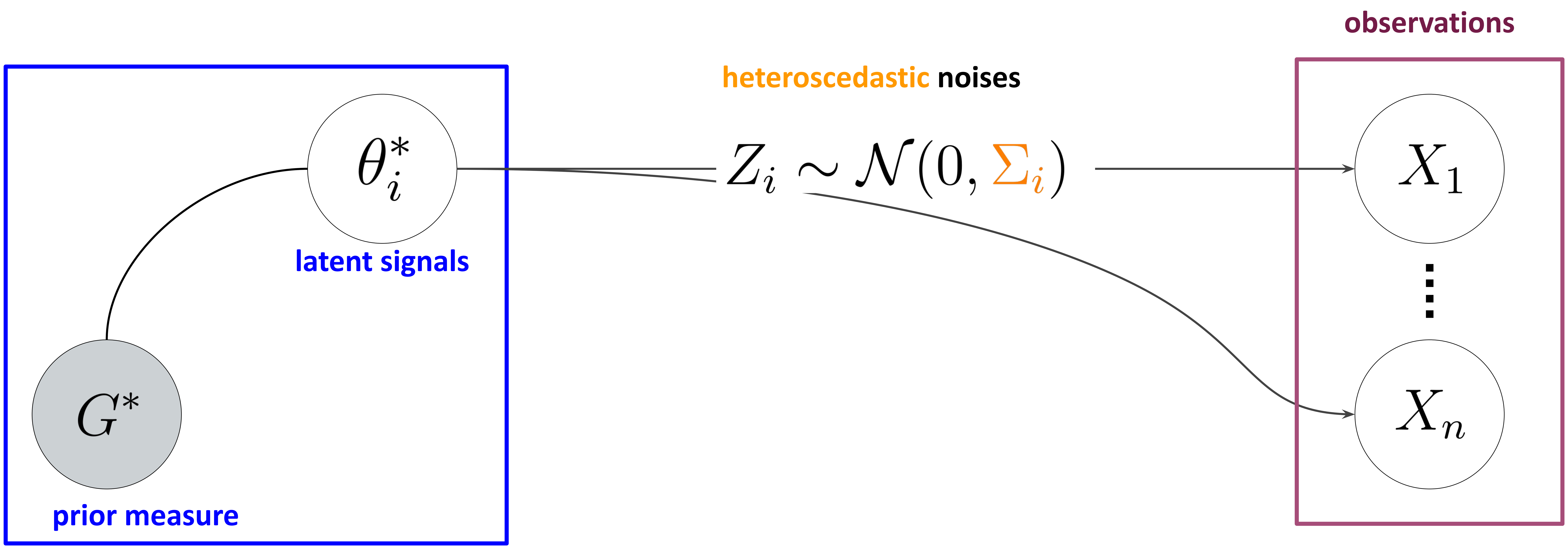
Score-based Diffusion Modeling for Empirical Bayes in Heteroscedastic Gaussian Mixtures

Gongyu Chen and **Ying Cui**
Department of IEOR, UC Berkeley

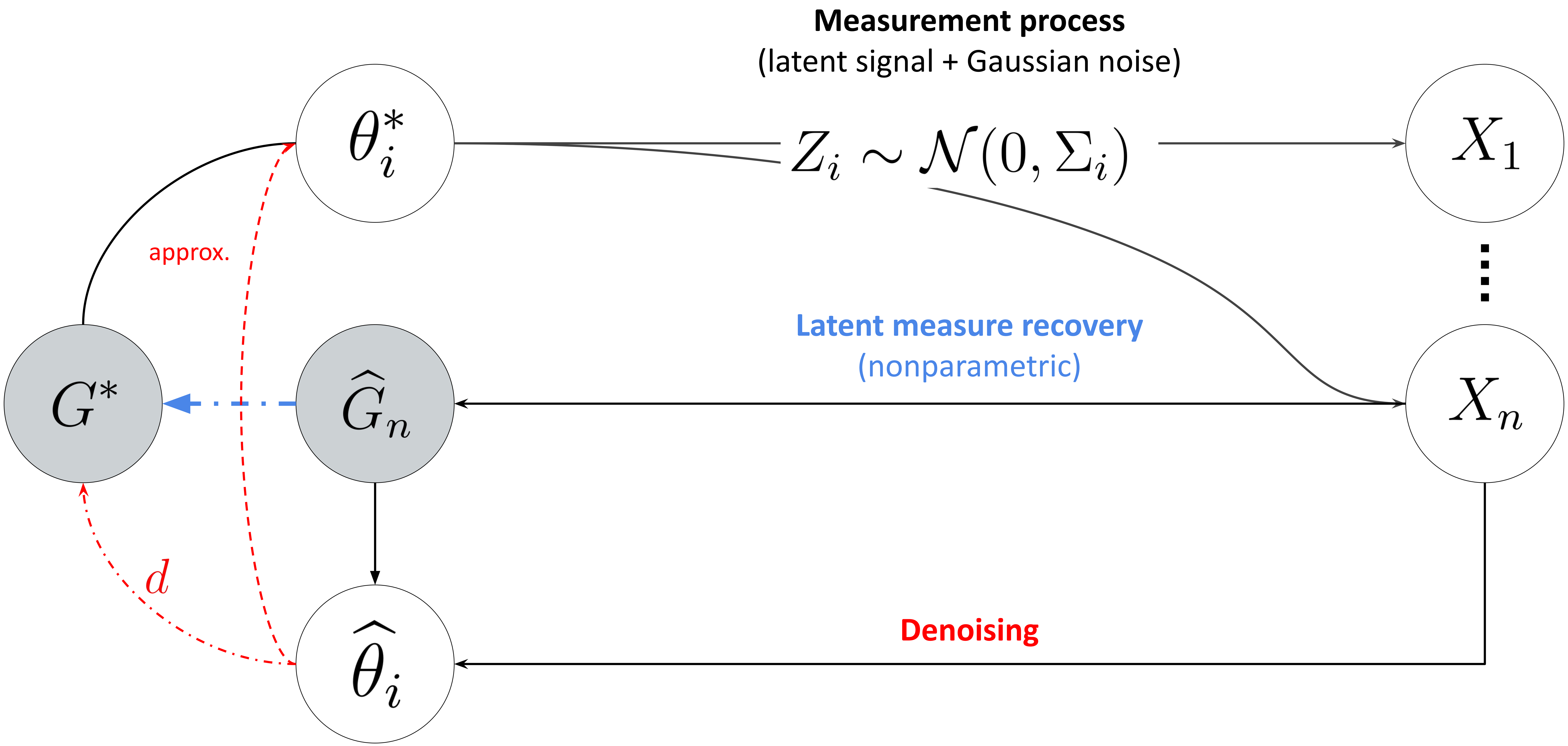
Gaussian Location Mixtures and Empirical Bayes

Measurement process
(latent signal + Gaussian noise)

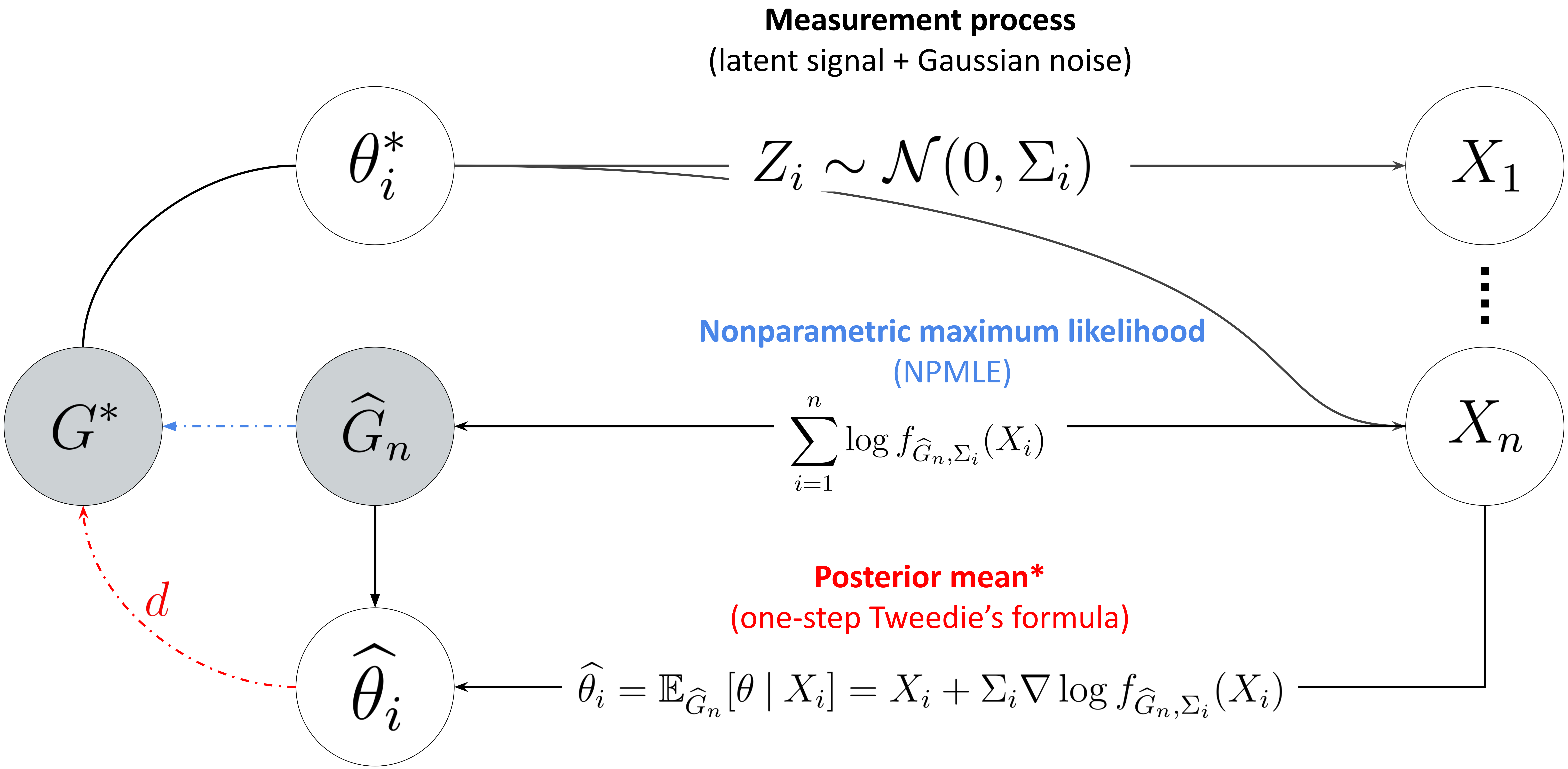
$$X_i = \theta_i^* + Z_i$$



Gaussian Location Mixtures and Empirical Bayes

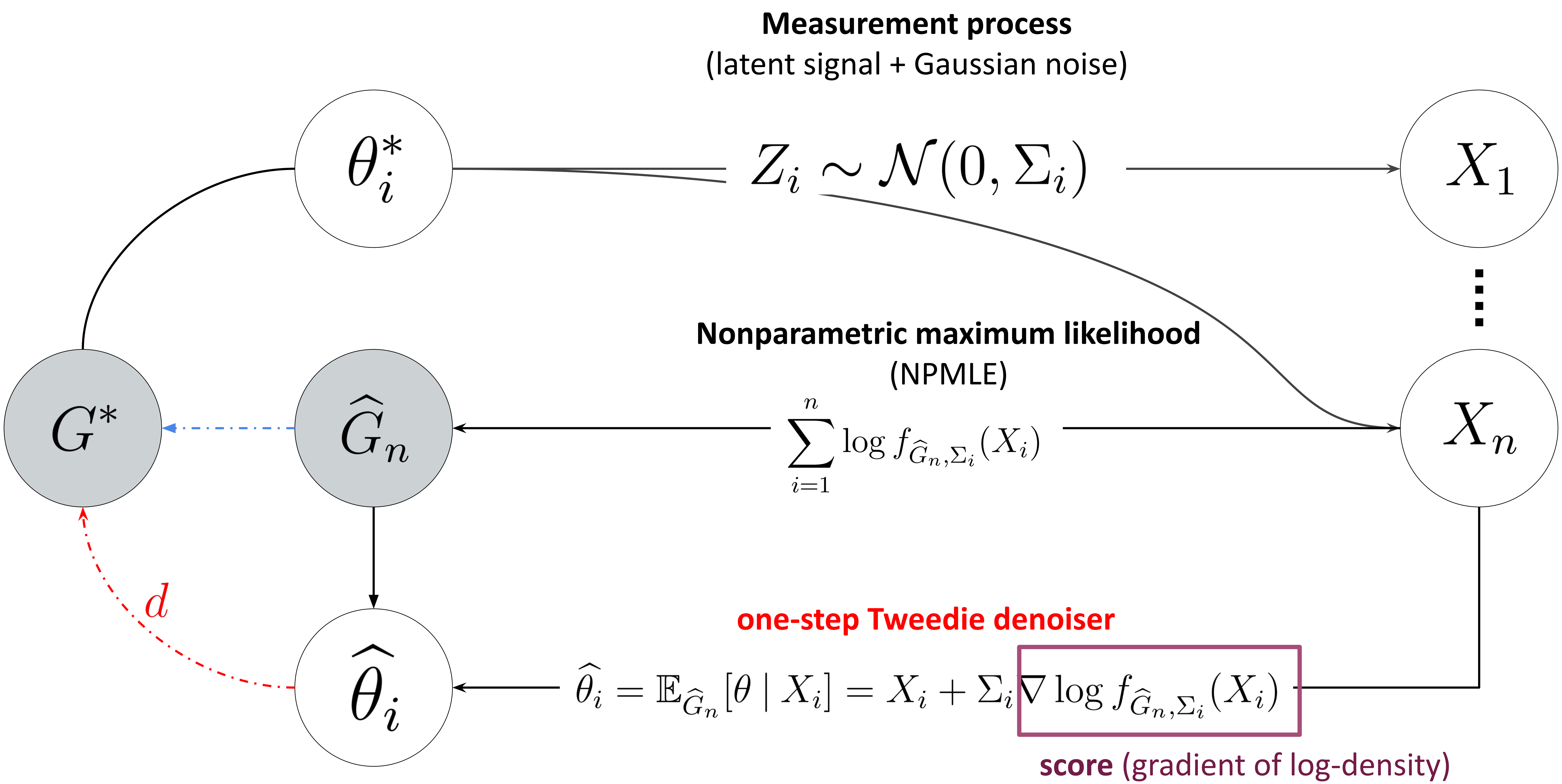


Classical EB estimators: NPMLE + one-step Tweedie



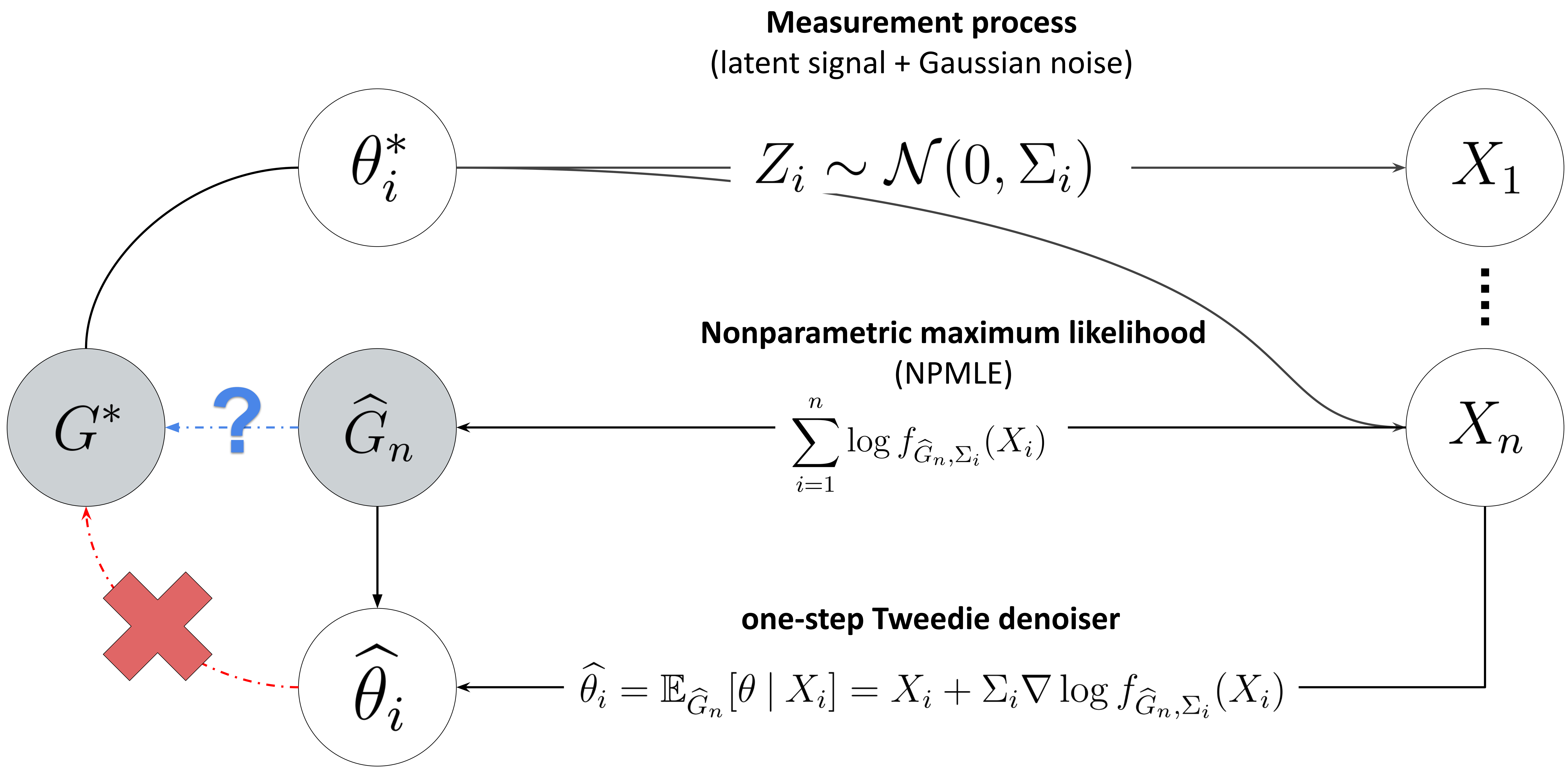
: <u>oracle</u> estimator (known G) minimizes the *Bayes risk*: $\widehat{\theta}_i^* := \mathbb{E}_{G^*}[\theta \mid X_i] = \arg\min_{T_i} \quad \mathbb{E}_{G^*}\|T_i(X_i) - \theta_i^*\|^2$

Classical EB estimators: NPMLE + one-step Tweedie



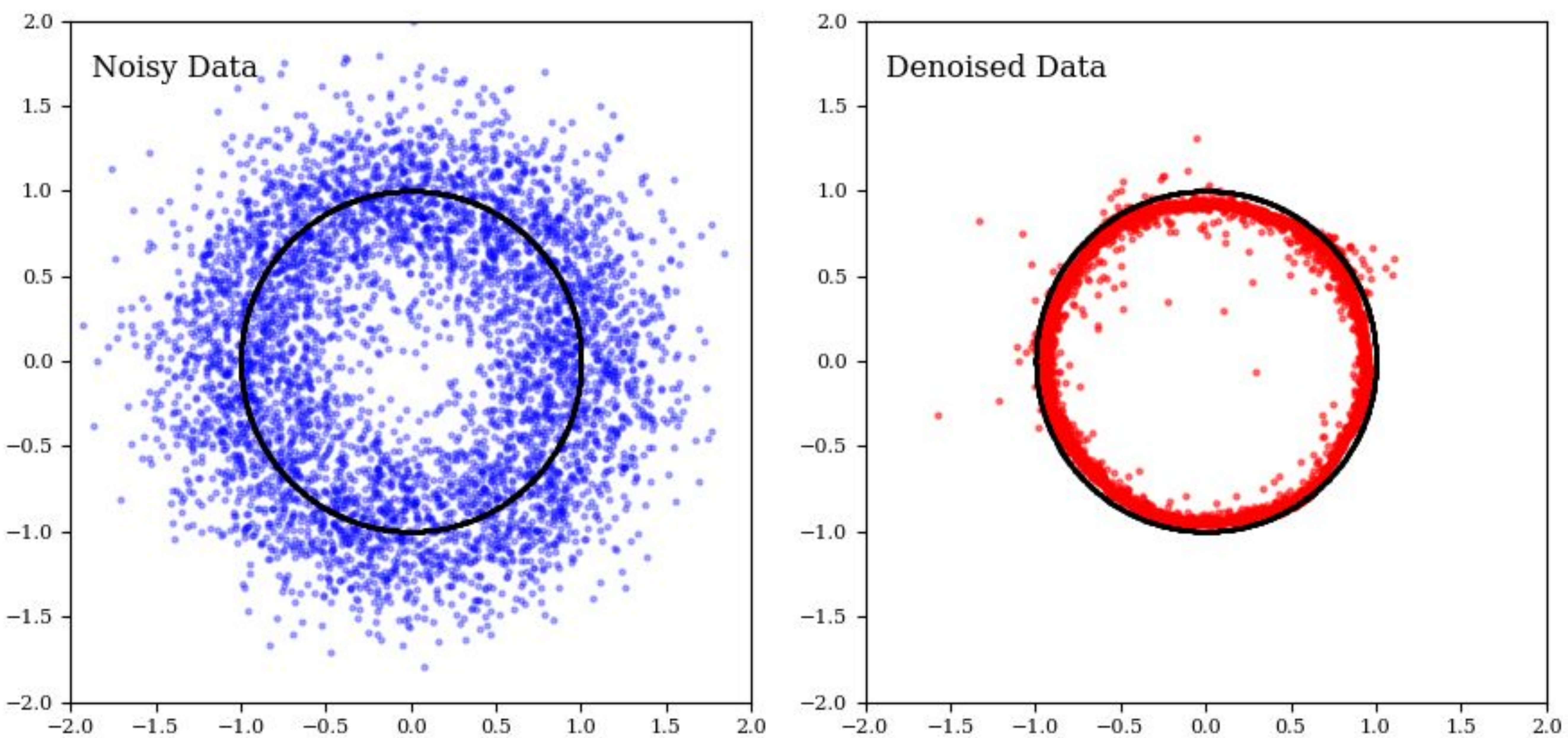
Issues with classical approach: over-shrinkage + computation in high-d

✓ **Theoretically,** NPMLE achieves *mini-max near-parametric* sample complexity rate* $O\big(n^{-1}(\log n)^{O(d)}\big)$



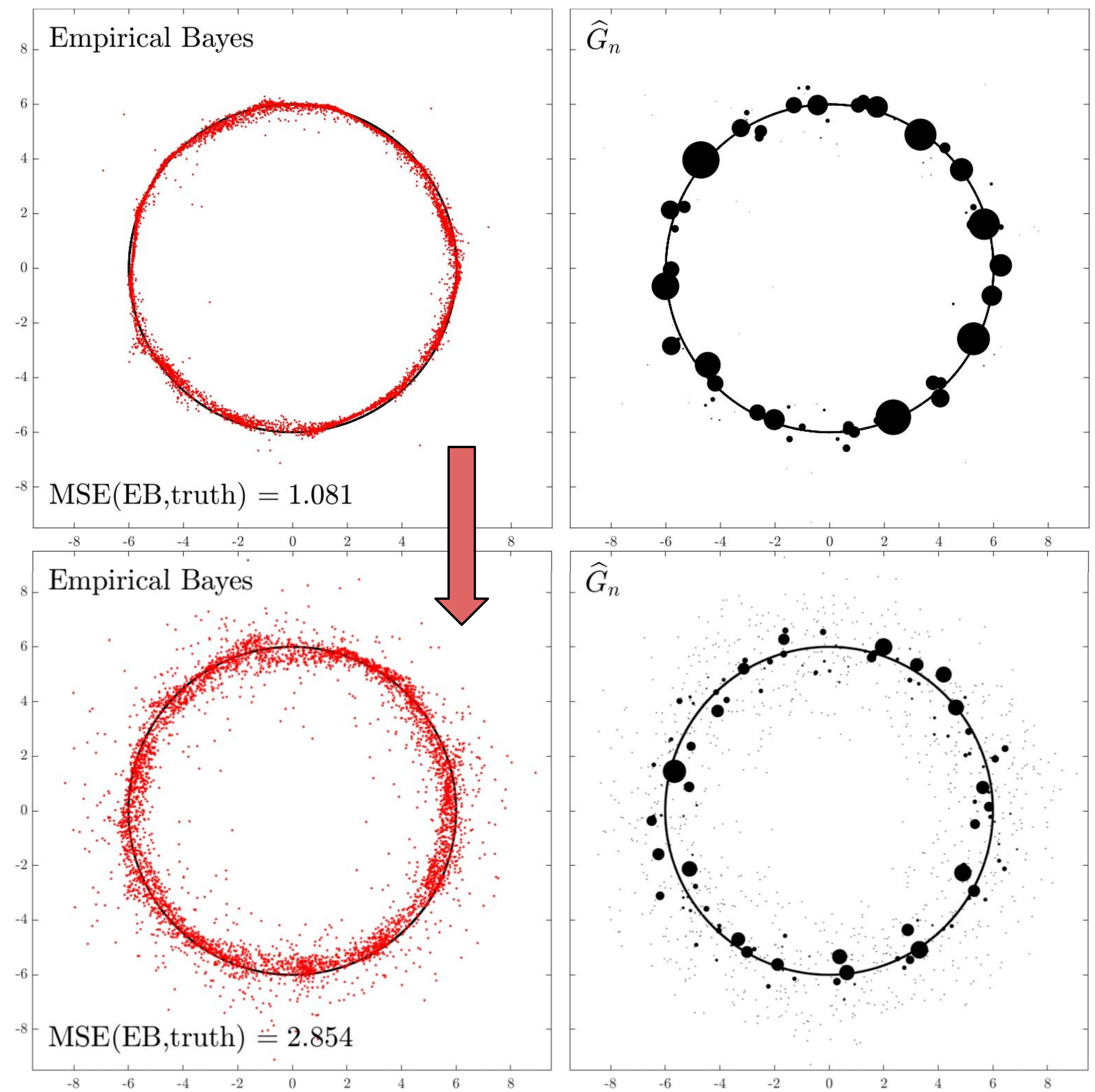
*: see [Soloff et al. 2025, Kim and Guntuboyina, 2022]

Over-shrinkage



- ***Denoised data*** *by Tweedie are over-shrunk relative to the* ***true unit circle****!*
- *Even the <u>oracle estimator</u> suffers from this systematic bias!*

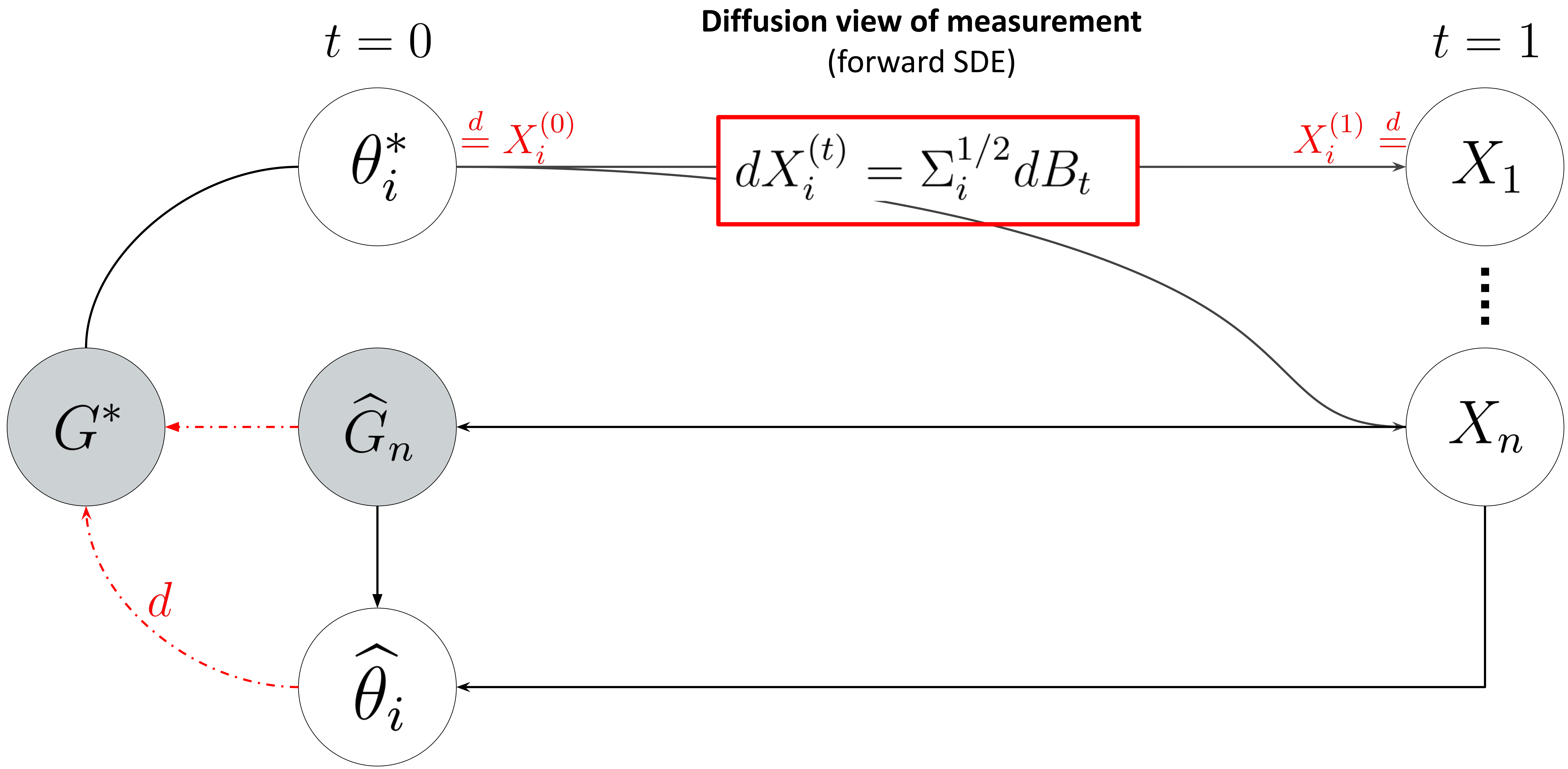
SOTA solver of NPMLE*



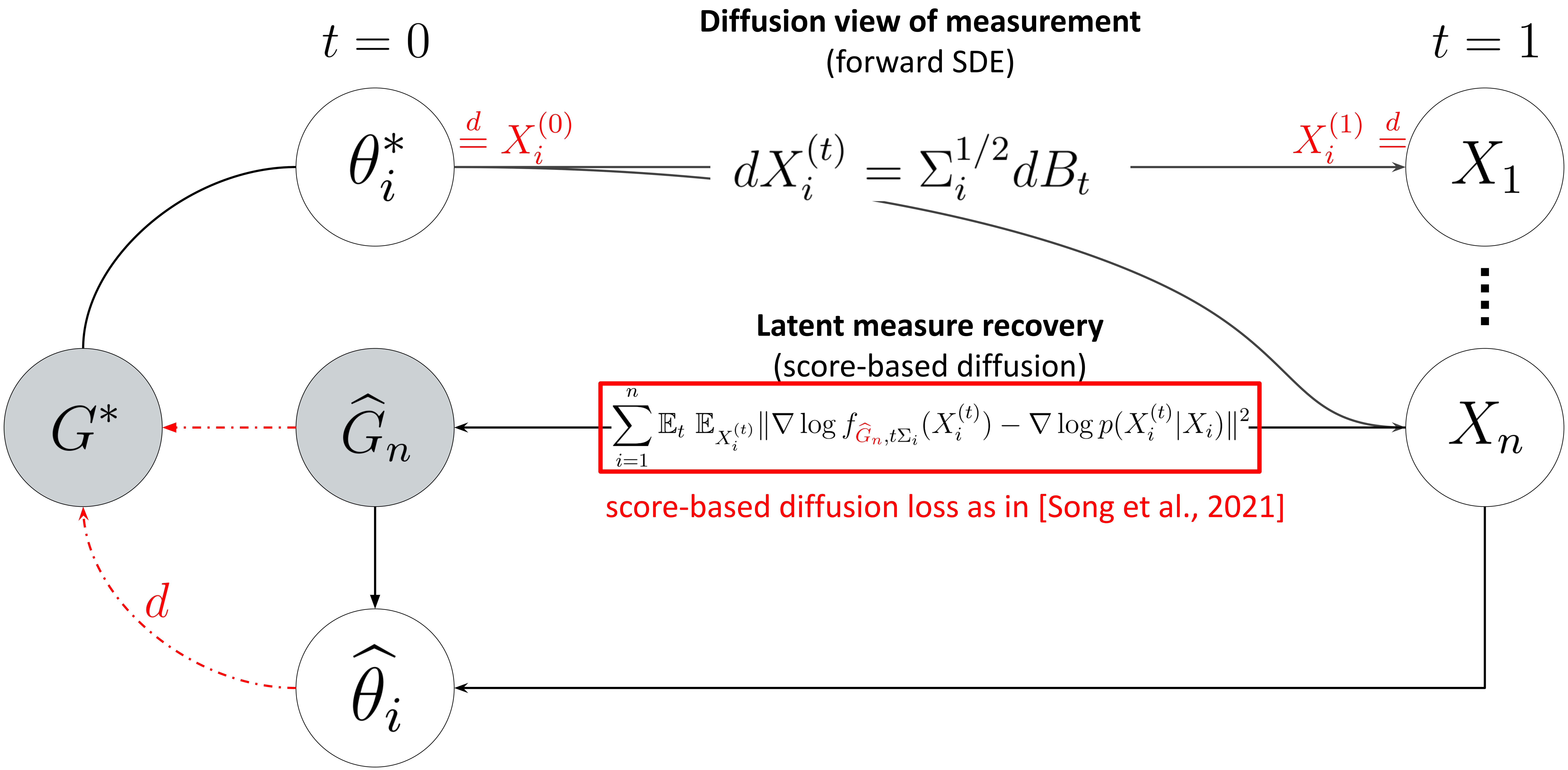
Quality of ***denoisers*** *deteriorates fast as from* ***d=3*** *to* ***d=9****!*

*: from [Zhang et al. 2024]

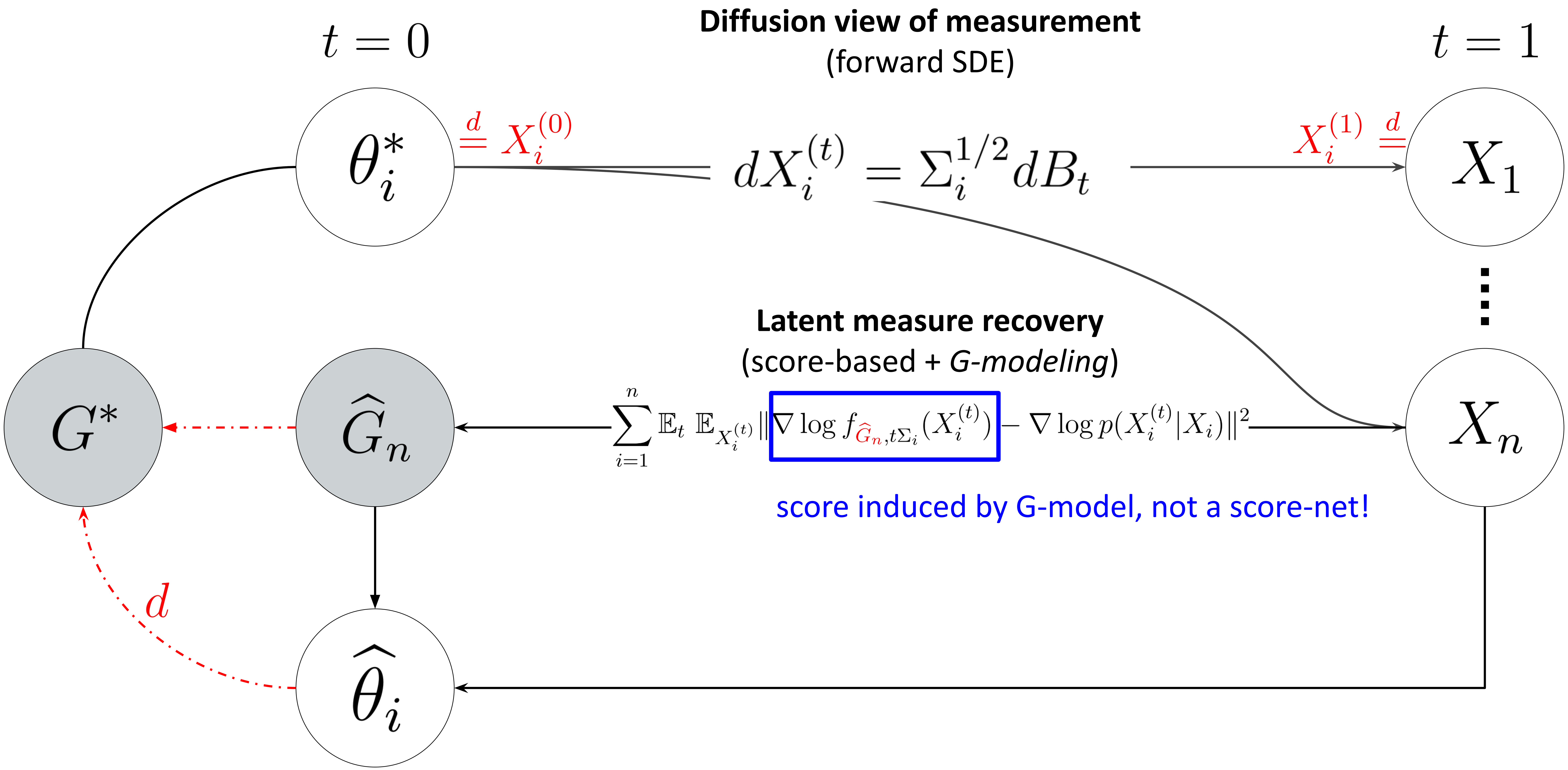
Our approach: Diffusion + Empirical Bayes G-modeling



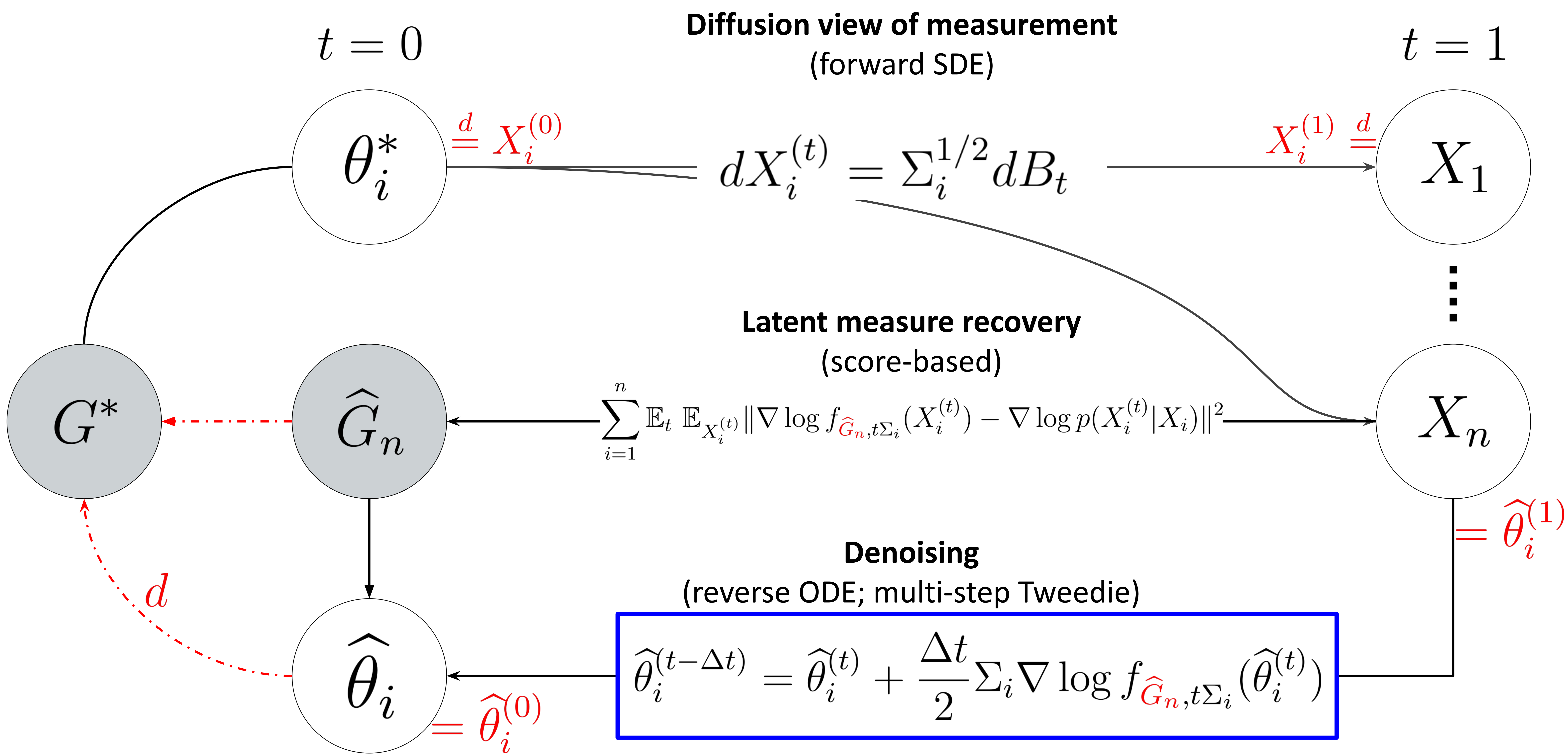
Our approach: Diffusion + Empirical Bayes G-modeling



Our approach: Diffusion + Empirical Bayes G-modeling



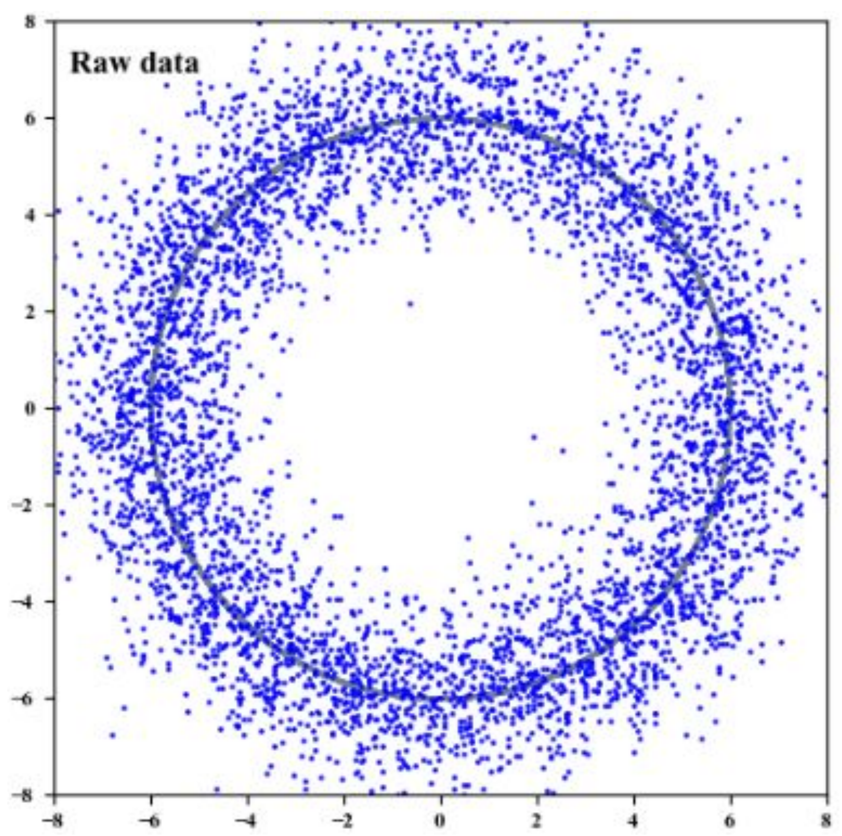
Our approach: Diffusion + Empirical Bayes G-modeling



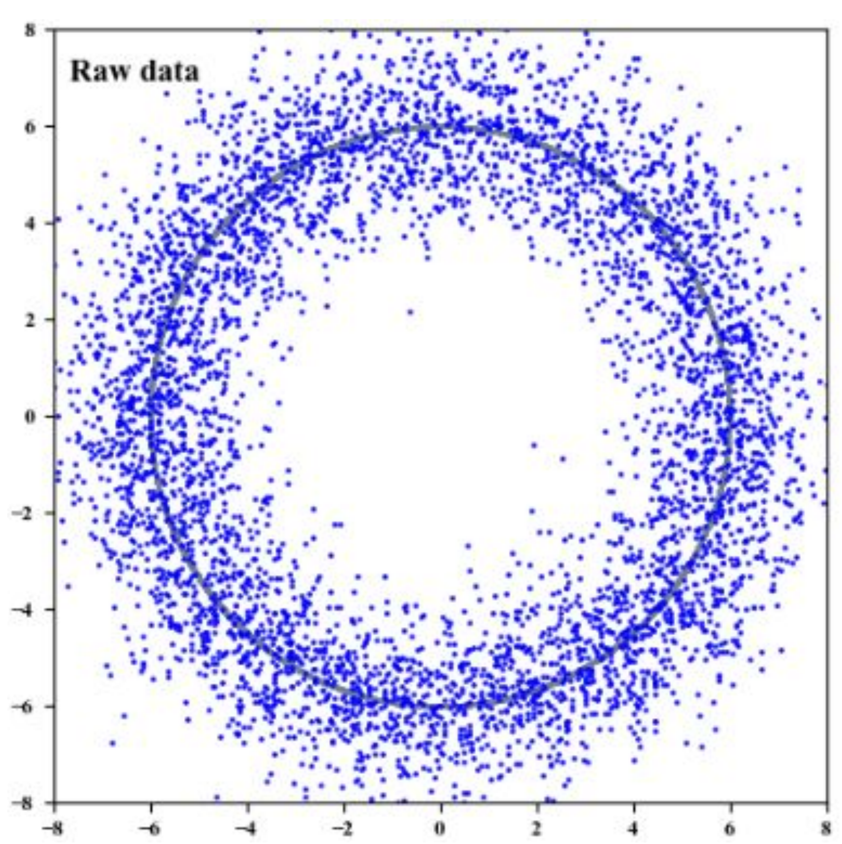
- When Δt=1, this recovers one-step Tweedie!
- When Ĝ=G*, this has no over-shrinkage bias!

Experiment results

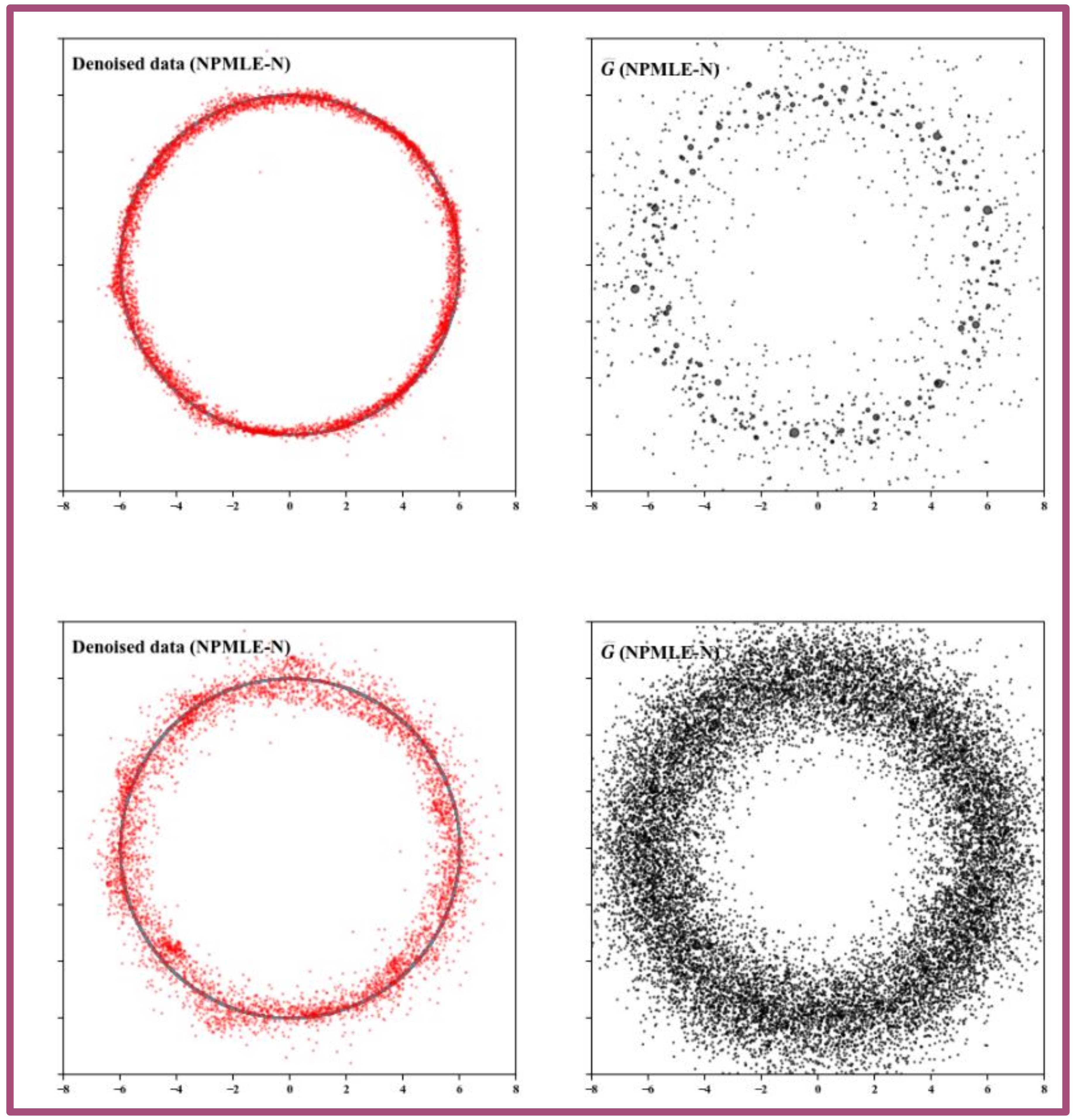
d=8



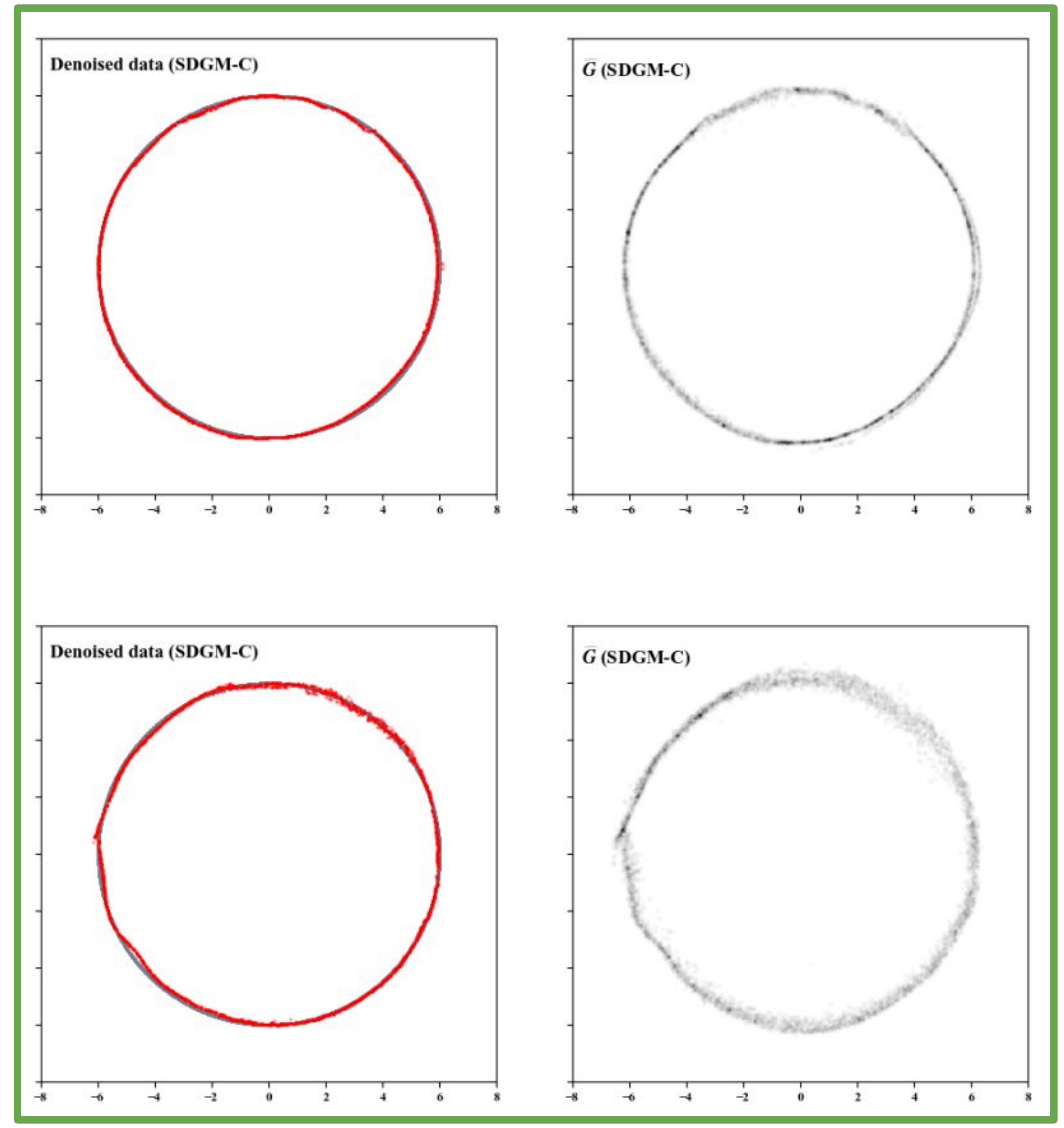
d=16



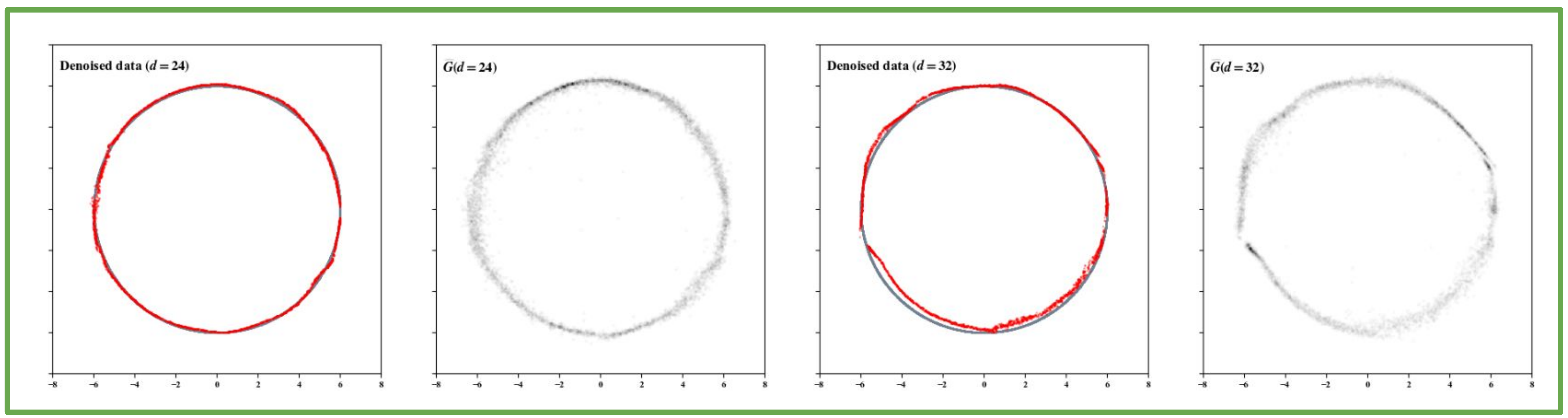
SOTA NPMLE solver [Zhang et al., 2024].



Our score-based solver [Chen and Cui, 2025]



Ours stays efficient in higher d = 24/32!



Theoretical guarantees

Near-parametric sample complexity on score estimation error!

Theorem 3. *Provided conditions in Assumption 2, let (t_0, T) satisfy $t_0 \leq T$ and*

$$\frac{1}{\log n} \leq \int_1^{t_0} g_i^2(t)dt \leq \int_1^T g_i^2(t)dt \leq 1.$$

Let $\widehat{G}_n$ be an optimal solution to Objective (11) constrained over the measure class $\mathcal{P}([-M, M]^d)$, then provided n sufficiently large, with probability at least $1 - n^{-2}$,

- *for score estimation:*

$$\begin{aligned}\mathbb{E}\overline{\mathfrak{F}}_{[t_0,T]}(q_{G^*}^{(t)} \| q_{\widehat{G}_n}^{(t)}) :=& \int_{t_0}^T \mathbb{E}_{x\sim q_{G^*}^{(t)}} \|\nabla \log q_{G^*}^{(t)}(x) - \nabla \log q_{\widehat{G}_n}^{(t)}(x)\|_2^2 dt \\ \leq& C_{d,M,(\overline{\sigma},\underline{\sigma}),(\underline{g},\overline{g})} \boxed{\frac{1}{n}(\log n)^{2d+3}};\end{aligned}$$

- *for density estimation at $t = t_0$:*

$$\mathbb{E}\mathfrak{H}^2(q_{G^*}^{(t_0)} \| q_{\widehat{G}_n}^{(t_0)}) \leq C'_{d,M,(\overline{\sigma},\underline{\sigma}),(\underline{g},\overline{g})} \frac{1}{n}(\log n)^{2d+3};$$

- *for the deconvolution risk,*

$$\mathbb{E}W_2^2\left(\frac{1}{n}\sum_{i=1}^n (\Sigma_i)^{-1/2}\#G^*, \frac{1}{n}\sum_{i=1}^n (\Sigma_i)^{-1/2}\#\widehat{G}_n\right) \lesssim \frac{1}{\log n}.$$

References

Chen, G., & Cui, Y. (2025). Score-Based Diffusion Modeling for Nonparametric Empirical Bayes in Heteroscedastic Gaussian Mixtures. To appear in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Kim, A. K., & Guntuboyina, A. (2022). Minimax bounds for estimating multivariate Gaussian location mixtures. *Electronic Journal of Statistics*, *16*(1), 1461-1484.

Soloff, J. A., Guntuboyina, A., & Sen, B. (2025). Multivariate, heteroscedastic empirical Bayes via nonparametric maximum likelihood. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *87*(1), 1-32.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021). Score-Based Generative Modeling through Stochastic Differential Equations. *International Conference on Learning Representations*.

Zhang, Y., Cui, Y., Sen, B., & Toh, K. C. (2024). On efficient and scalable computation of the nonparametric maximum likelihood estimator in mixture models. *Journal of Machine Learning Research*, *25*(8), 1-46.