Proximal-Based Generative Modeling for Bayesian Inverse Problems

Boyang Zhang[1] Zhiguo Wang[2] Ya-Feng Liu[3]

[1]School of Advanced Interdisciplinary Sciences,
University of Chinese Academy of Sciences

[2]School of Mathematics,
Sichuan University

[3]School of Mathematical Sciences,
Beijing University of Posts and Telecommunications

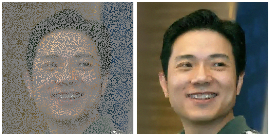
Forty-Third International Conference on Machine Learning

Inverse Problems

- Given measurement $\boldsymbol{y} \in \mathbb{R}^m$ of the form

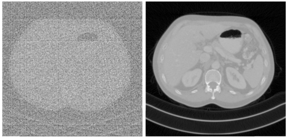
$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{x}) + \boldsymbol{\xi},$$

  where $\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^m$ denotes a forward operator, $\boldsymbol{x} \in \mathbb{R}^d$ is an unknown signal, and $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma_\xi^2 I)$ is the additive noise.
- Ill-Posed: $m \ll d$ (fewer measurements than pixels)
- Many applications: Super resolution, inpainting, deblurring ...



Random Inpainting

GaussianDeblurring



CT Reconstruction

Optimization-Based Methods

- Intuition: reconstruct the unknown signal by identifying the most likely one under the posterior distribution.

$$\min_{\boldsymbol{x}} \ \beta\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{x})\|^2 + g(\boldsymbol{x}),$$

where g is a prior, β is a weighting coefficient.

- Leading to: maximum a posteriori (MAP) estimator.
- Algorithms: ADMM, FISTA, ...

- Pros: provide a good point estimate, efficiency.
- Cons: fail to reflect **uncertainty** or handle **unknown prior**.

Sampling-Based Methods

- Bayesian formulation: sample from posterior

$$\pi(\boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}) \propto e^{-U(\boldsymbol{x})}, \quad U(\boldsymbol{x}) = \beta f_{\boldsymbol{y}}(\boldsymbol{x}) + g(\boldsymbol{x}),$$

  where $f_{\boldsymbol{y}}(\boldsymbol{x}) = \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{x})\|^2$ is the data fidelity (convex), g is a given prior (nonsmooth).
- Algorithms: proximal Langevin, subgradient Langevin, ...

- Pros: approximates complex distributions, global convergence.
- Cons: time-consuming, cannot handle **unknown prior**.

Sampling from Unknown Distribution: Diffusion Models

- Forward SDE:

$$d\boldsymbol{x}_t^{\rightarrow} = a(t)\boldsymbol{x}_t^{\rightarrow} dt + b(t) d\boldsymbol{B}_t.$$

- Reverse SDE:

$$d\boldsymbol{x}_t^{\leftarrow} = \left\{a(t)\boldsymbol{x}_t^{\leftarrow} - b^2(t)\nabla \log p_t(\boldsymbol{x}_t^{\leftarrow})\right\} dt + b(t) d\bar{\boldsymbol{B}}_t.$$

- Denoising score matching:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathbb{E}\left[\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \nabla \log p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)\|_2^2\right].$$

- After training, the reverse SDE is simulated with standard numerical solvers to generate samples.

Diffusion Models for Bayesian Inverse Problems

- Conditional reverse SDE:

$$d\boldsymbol{x}_t = \left\{a(t)\boldsymbol{x}_t - b^2(t)\nabla \log p_t(\boldsymbol{x}_t|\boldsymbol{y})\right\}dt + b(t)d\bar{\boldsymbol{B}}_t,$$

where the time-dependent conditional score

$$\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t|\boldsymbol{y}) = \underbrace{\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t)}_{\text{prior score}} + \underbrace{\nabla_{\boldsymbol{x}} \log p(\boldsymbol{y}|\boldsymbol{x}_t)}_{\text{log-likelihood}}.$$

- **Main technical challenge**: the time-dependent likelihood term

$$\nabla_{\boldsymbol{x}} \log p(\boldsymbol{y}|\boldsymbol{x}_t) = \int_{\mathbb{R}^d} p(\boldsymbol{y}|\boldsymbol{x}_0)p(\boldsymbol{x}_0|\boldsymbol{x}_t)d\boldsymbol{x}_0.$$

Existing Works: DPS Family

- Goal: approximate the likelihood term by estimating $p(\boldsymbol{x}_0|\boldsymbol{x}_t)$:

$$\nabla_{\boldsymbol{x}} \log p(\boldsymbol{y}|\boldsymbol{x}_t) = \int_{\mathbb{R}^d} p(\boldsymbol{y}|\boldsymbol{x}_0)p(\boldsymbol{x}_0|\boldsymbol{x}_t)d\boldsymbol{x}_0.$$

- Given denoiser $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t) = (\boldsymbol{x}_t + (1-\bar{\alpha}_t)\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t))/\sqrt{\bar{\alpha}_t}$:
  1. DPS with Laplace approximation of $p(\boldsymbol{x}_0|\boldsymbol{x}_t)$:

  $$p_t(\boldsymbol{y}|\boldsymbol{x}_t) \sim \mathcal{N}(A\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t), \sigma_\xi^2 I).$$

  2. ΠGDM:

  $$p_t(\boldsymbol{y}|\boldsymbol{x}_t) \sim \mathcal{N}(A\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t), \sigma_\xi^2 I + r_t^2 AA^\top).$$

- Cons: calculating $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{y}|\boldsymbol{x}_t)$ requires **backpropagation** through network $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t)$, which is computationally demanding and memory intensive.

Existing Works: Plug-and-Play

- Goal: adapt convex optimization by substituting classical filters with diffusion models acting as regularizing denoisers.
- Examples:
  1. DiffPIR with half-quadratic splitting:

$$\min_{\boldsymbol{x},\boldsymbol{z}} \; f(\boldsymbol{x}) + g(\boldsymbol{z}) + \frac{1}{2\lambda/\mu}\|\boldsymbol{x}-\boldsymbol{z}\|^2.$$

The subproblem uses first-order approximation of prior:

$$\boldsymbol{z}_k = \arg\min_{\boldsymbol{z}} \|\boldsymbol{z}-\boldsymbol{x}_k\|^2/(2\lambda) + g(\boldsymbol{z}),$$

where $g(\boldsymbol{z}) \approx g(\boldsymbol{x}_k) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_k, t_k)^\top(\boldsymbol{z}-\boldsymbol{x}_k)$.

- Cons: based on traditional Stein score matching and first-order approximation, cannot handle **nonsmooth priors**.

Key Research Question

1. DPS Family requires **backpropagation** through network $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t)$;
2. PnP methods cannot handle **nonsmooth prior** $p(\boldsymbol{x}_0) \propto e^{-g(\boldsymbol{x}_0)}$.

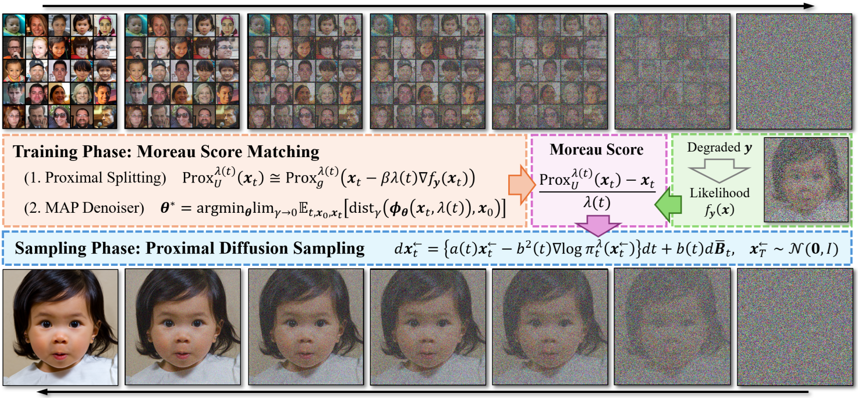
Can we construct an **alternative score function** that replaces the time-dependent posterior score while still guaranteeing correct sampling from the posterior distribution?

Our answer: Proximal-Based Generative Modeling (PGM)

Main Contributions

- **Novel framework bridging optimization and diffusion:**
  1. Establish Moreau–Yosida-Gaussian equivalence.
  2. Introduce closed-form Moreau score via proximal operators.
- **Unsupervised Moreau score matching:**
  1. Learn proximal operator from prior samples only.
  2. Enables plug-and-play posterior sampling.
- **Improved theory and experiments:**
  1. Non-asymptotic convergence guarantees.
  2. State-of-the-art results in generation quality and sampling time.

PGM Framework



Motivation: Moreau–Yosida–Gaussian Equivalence

- Consider the VE-SDE forward process:

$$d\boldsymbol{x}_t = b(t)d\boldsymbol{B}_t.$$

- Reverse SDE of diffusion model:

$$d\boldsymbol{x}_t = \Big\{ -b^2(t)\nabla \log p_t(\boldsymbol{x}_t^{\leftarrow}) \Big\} dt + b(t)d\bar{\boldsymbol{B}}_t.$$

- Marginal density:

$$\pi_t(\boldsymbol{x}_t) \propto \int_{\mathbb{R}^d} \pi(\boldsymbol{x}_0) \exp\left\{ -\frac{\|\boldsymbol{x}_0 - \boldsymbol{x}_t\|^2}{2\lambda(t)} \right\} d\boldsymbol{x}_0,$$

where $\lambda(t) = \int_0^t b^2(s)ds$.

Motivation: Moreau–Yosida–Gaussian Equivalence

Lemma 1 (Quadratic Form)

Consider $\pi(\boldsymbol{x}_0) \propto \exp\{-U(\boldsymbol{x}_0)\}$ *with positive semidefinite quadratic potential* U *and marginal density*

$$\pi_t(\boldsymbol{x}_t) \propto \int_{\mathbb{R}^d} \pi(\boldsymbol{x}_0) \exp\left\{-\frac{\|\boldsymbol{x}_0 - \boldsymbol{x}_t\|^2}{2\lambda(t)}\right\} d\boldsymbol{x}_0.$$

Define the Moreau approximation

$$\pi_t^\lambda(\boldsymbol{x}_t) \propto \sup_{\boldsymbol{x}_0 \in \mathbb{R}^d} \left\{\pi(\boldsymbol{x}_0) \exp\left\{-\frac{\|\boldsymbol{x}_0 - \boldsymbol{x}_t\|^2}{2\lambda(t)}\right\}\right\}.$$

Then we have

$$\pi_t^\lambda(\boldsymbol{x}_t) \propto \pi_t(\boldsymbol{x}_t).$$

Moreau Score: Closed-Form via Proximal Operator

- Proposed reverse SDE using Moreau score:

$$d\boldsymbol{x}_t = \big\{ -b^2(t)\nabla \log \pi_t^{\lambda}(\boldsymbol{x}_t)\big\}dt + b(t)d\bar{\boldsymbol{B}}_t.$$

- Moreau score admits closed form:

$$\nabla \log \pi_t^{\lambda}(\boldsymbol{x}_t) = \frac{\mathrm{Prox}_U^{\lambda(t)}(\boldsymbol{x}_t) - \boldsymbol{x}_t}{\lambda(t)}.$$

- By Tweedie's formula:

$$\nabla_{\boldsymbol{x}} \log \pi_t(\boldsymbol{x}_t) = \frac{\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t] - \boldsymbol{x}_t}{\lambda(t)}.$$

- Difference: the discrepancy between the **posterior mean** $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]$ and the **posterior mode** $\mathrm{Prox}_U^{\lambda(t)}(\boldsymbol{x}_t)$.

Connection between Moreau Score and Stein Score

Proposition 2 (Discrepancy)

1. *If the posterior* $p(\boldsymbol{x}_0|\boldsymbol{x}_t, \boldsymbol{y})$ *is symmetric unimodal, then*

$$\nabla \log \pi_t(\boldsymbol{x}_t) = \nabla \log \pi_t^{\lambda}(\boldsymbol{x}_t).$$

2. *If* $f_{\boldsymbol{y}}$ *is convex, then*

$$\left\| \nabla \log \pi_t(\boldsymbol{x}_t) - \nabla \log \pi_t^{\lambda}(\boldsymbol{x}_t) \right\| \leq \sqrt{\frac{2d}{\lambda(t)}}.$$

In particular, if $f_{\boldsymbol{y}}$ *is* m*-strongly convex, then*

$$\left\| \nabla \log \pi_t(\boldsymbol{x}_t) - \nabla \log \pi_t^{\lambda}(\boldsymbol{x}_t) \right\| \leq \sqrt{\frac{2d}{\beta m \lambda^2(t) + \lambda(t)}}.$$

Moreau Score Matching

- Require: $\mathrm{Prox}_U^{\lambda(t)}(\boldsymbol{x}_t)$ for $U(\boldsymbol{x}) = \beta f_{\boldsymbol{y}}(\boldsymbol{x}) + g(\boldsymbol{x})$.
- Use proximal splitting:

$$\mathrm{Prox}_U^{\lambda(t)}(\boldsymbol{x}_t) \approx \mathrm{Prox}_g^{\lambda(t)}\left(\boldsymbol{x}_t - \beta\lambda(t)\nabla f_{\boldsymbol{y}}(\boldsymbol{x}_t)\right),$$

  where $f_{\boldsymbol{y}}$ is explicitly given, g is unknown.
- Intuition: the proximal operator is equivalent to the MAP denoiser given Gaussian corrupted observation $\boldsymbol{x}_t$:

$$\mathrm{Prox}_g^{\lambda(t)}(\boldsymbol{x}_t) = \underset{\boldsymbol{x}_0 \in \mathbb{R}^d}{\arg\max}\ p(\boldsymbol{x}_0|\boldsymbol{x}_t).$$

- We use proximal network $\phi_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \lambda)$ to approximate $\mathrm{Prox}_g^{\lambda}(\boldsymbol{x}_t)$.

Learning the Proximal Operator from Data

Proposition 3 (Proximal Matching)

Let $\boldsymbol{x}_t = \boldsymbol{x}_0 + \sqrt{\lambda(t)}\,\boldsymbol{\xi}_t$. *Consider* $\boldsymbol{\theta}^*$ *that minimizes the Bayes risk:*

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \lim_{\zeta \to 0} \mathbb{E}\left[\mathit{dist}_\zeta\left(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \lambda(t)), \boldsymbol{x}_0\right)\right],$$

where the expectation is taken over $t \sim \mathcal{U}[0, T]$, $\boldsymbol{x}_0 \sim \exp\{-g(\boldsymbol{x}_0)\}$, $\boldsymbol{x}_t \sim p(\boldsymbol{x}_t|\boldsymbol{x}_0)$, *and* $\mathit{dist}_\zeta(\boldsymbol{x}, \boldsymbol{x}') = 1 - \mathcal{N}(\boldsymbol{x} - \boldsymbol{x}'; \mathbf{0}, \zeta^2 I)$. *Then*

$$\boldsymbol{\phi}_{\boldsymbol{\theta}^*}(\boldsymbol{x}_t, \lambda(t)) = \arg\max_{\boldsymbol{x}_0 \in \mathbb{R}^d}\; p(\boldsymbol{x}_0|\boldsymbol{x}_t) = \mathrm{Prox}_g^{\lambda(t)}(\boldsymbol{x}_t).$$

- Intuition: as $\zeta \to 0$, the Gaussian kernel $\mathcal{N}(\cdot; \mathbf{0}, \zeta^2 I)$ behaves like a **Dirac delta**, which forces ϕ to output the posterior mode.

Reverse Process with Learned Proximal Operator

- Consider a general reverse process

$$d\boldsymbol{x}_t^{\leftarrow} = \left\{ a(t)\boldsymbol{x}_t^{\leftarrow} - b^2(t)\nabla \log \pi_t^{\lambda}(\boldsymbol{x}_t^{\leftarrow}) \right\} dt + b(t)d\bar{\boldsymbol{B}}_t.$$

- Generalized Moreau score:

$$\nabla \log \pi_t^{\lambda}(\boldsymbol{x}_t) = \frac{\mu(t)\,\mathrm{Prox}_U^{\lambda(t)}\left(\frac{\boldsymbol{x}_t}{\mu(t)}\right) - \boldsymbol{x}_t}{\sigma^2(t)}.$$

- Discretization of reverse process via exponential integrator:

$$\bar{\boldsymbol{x}}_{k+1} = \alpha_{1,k}\bar{\boldsymbol{x}}_k + \alpha_{2,k}\boldsymbol{P}_k + \alpha_{3,k}\boldsymbol{\xi}_k, \quad k = 1, \ldots, K$$

where $\boldsymbol{P}_k = \boldsymbol{\phi}_{\boldsymbol{\theta}^*}\left(\frac{\bar{\boldsymbol{x}}_k}{\mu(\tau_k)} - \beta\lambda(\tau_k)\nabla f\left(\frac{\bar{\boldsymbol{x}}_k}{\mu(\tau_k)}\right), \lambda(\tau_k)\right)$.

Algorithm 1 Proximal-Based Generative Modeling

Require: Realizations $\{\boldsymbol{x}^{(i)}\}_{i=1}^{N_s} \sim \exp\{-g\}$, objective $f_{\boldsymbol{y}}$, inverse temperature β, learning schedule $\{\zeta_{\text{Epoch}}\}$.

Ensure: Sample $\boldsymbol{x} \sim \exp\{-[\beta f_{\boldsymbol{y}}(\boldsymbol{x}) + g(\boldsymbol{x})]\}$.

(Training Phase: Moreau Score Matching)
for Epoch $= 1$ to MaxEpoch **do**
    Sample $\boldsymbol{x}_0 \sim \exp\{-g(\boldsymbol{x})\}, t \in \mathcal{U}[0, T], \boldsymbol{x}_t|\boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}, \lambda(t)I)$.
    Take stochastic gradient descent on $\phi_{\boldsymbol{\theta}}(\boldsymbol{x}, \lambda)$.
end for
(Sampling Phase: Proximal Diffusion Sampling)
Initialize $\bar{\boldsymbol{x}}_0 \sim p_T$.
for $k = 0$ to K **do**
    Update $\bar{\boldsymbol{x}}_{k+1} = \alpha_{1,k}\bar{\boldsymbol{x}}_k + \alpha_{2,k}\boldsymbol{P}_k + \alpha_{3,k}\boldsymbol{\xi}_k$.
end for
Return Sample $\boldsymbol{x} = \bar{\boldsymbol{x}}_{K+1}$.

Main Assumption

Assumption 4

1. $f_{\boldsymbol{y}}$ *is proper, closed,* L*-smooth, and* m*-strongly convex.*
2. g *is convex with compact domain* $\mathcal{X}$*.*
3. $\mu(t) \in [0, 1]$ *is nonincreasing and* $m_\mu \leq |\nabla \log \mu(t)| \leq M_\mu \leq 1/2$*.*
4. $\lambda(t) \in [0, \bar{\lambda}]$ *is nondecreasing and* $1 \leq m_\lambda \leq |\nabla \log \lambda(t)| \leq M_\lambda$*.*
5. $\delta := \sup_k \{t_{k+1} - t_k\} < \infty$*.*

- Assumption (1-2) are wildly used in composite optimization (Nesterov, 2005; Ehrhardt, 2024) to control the iteration sequences;
- Assumption (3-4) control the growth of the coefficients and is satisfied by practical SDEs (e.g., VE-SDE and VP-SDE);
- Assumption (5) bounds the stepsize, which is standard for analyzing the convergence (Gen Li, 2024; Yuejie Chi, 2026).

Bound on Score Approximation

Theorem 5 (Score estimation error)

Consider the diffusion process $d\boldsymbol{x}_t^{\rightarrow} = a(t)\boldsymbol{x}_t^{\rightarrow}dt + b(t)d\boldsymbol{B}_t$. *Denote*

$$\boldsymbol{P}_t = \phi_{\boldsymbol{\theta}^*}\left(\frac{\boldsymbol{x}_t}{\mu(t)} - \beta\lambda(t)\nabla f_{\boldsymbol{y}}(\frac{\boldsymbol{x}_t}{\mu(t)}), \lambda(t)\right).$$

Then there exists $M \geq 0$ *such that*

$$\left\|\nabla \log \pi_t(\boldsymbol{x}_t) - \frac{\mu(t)\boldsymbol{P}_t - \boldsymbol{x}_t}{\sigma^2(t)}\right\| \leq \frac{M\left(1 + \|\boldsymbol{x}_t\|\right)}{\sigma^2(t)}.$$

- Error decays as β (inverse temperature) increases.
- Valid for nonsmooth priors via proximal operators.

Non-Asymptotic Convergence

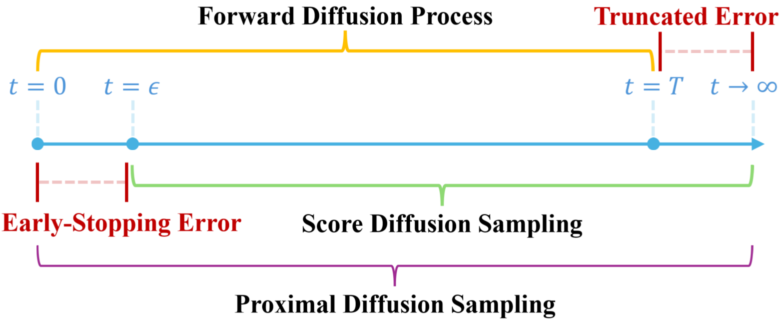
Theorem 6 (Wasserstein-1 convergence)

For sufficiently large T, there exist D_1, D_2, D_3 such that

$$\mathcal{W}_1\left(\mathrm{Law}(\bar{\boldsymbol{x}}_K), \pi\right) \leq D_1 \exp\left[-m_\mu T\right] + D_2 M + D_3 \delta^{1/2}.$$

- $\mathcal{W}_1\left(\mathrm{Law}(\bar{\boldsymbol{x}}_K), \pi\right)$ vanishes as $T \to \infty$, $M \to 0$, and $\delta \to 0$.

Sampling Error Decomposition



- Early-stopping error of order $\mathcal{O}(\sqrt{\epsilon})$ arises to avoid singularity of Stein score near $t = 0$ (De Bortoli, 2022; Gen Li, 2024).
- This is resolved by the well-defined Moreau score $\nabla \log \pi_t^{\lambda}(\boldsymbol{x}_t)$.

Toy Example: Quadratic Programming

- Optimization problem:

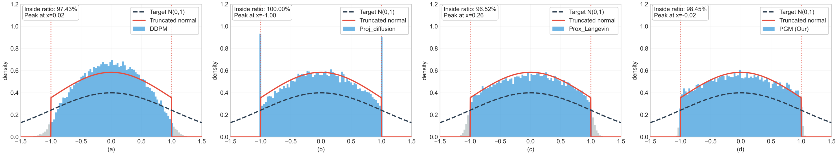
$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^d} \quad & f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top A\boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{x} \in \mathcal{X} \triangleq \{\boldsymbol{x} : \|x\|_2 \leq r\}, \end{aligned}$$

- Sampling task:

$$\pi(\boldsymbol{x}) \propto \exp\{-\beta f(\boldsymbol{x})\}\, \mathbb{I}_{\mathcal{X}}(\boldsymbol{x}).$$

- Two cases:
  1. Simple case: $\pi(\boldsymbol{x}) \propto \mathcal{N}(0,1)\mathbb{I}_{[-1,1]}$ where $A = I, \boldsymbol{b} = \mathbf{0}, r = 1$.
  2. General case: $A, \boldsymbol{b}, r$ are randomly generated, realizations are sampled uniformly from $\mathcal{X}$ for training.

Simple Case: Sampling from Truncated Normal



- Comparison: (a) DDPM, (b) projected diffusion model, (c) proximal Langevin, and (d) PGM.
- Score-based methods (a) and (b) fail to handle constraint. Proximal-based methods (c) and (d) perform better.
- PGM achieves better feasibility (**inside-ratio**$= 98.45\%$) and optimality (**peak at** $x = -0.02$).

General Case: Effect of # Iterations and Temperature

Metric	10000 PGM samples, $\beta = 10$				
	$K = 0$	$K = 1$	$K = 5$	$K = 10$	$K = 20$
Feasibility	0.00%	7.43%	97.43%	99.61%	99.73%

Metric	10000 PGM samples, $K = 10$				
	$\beta = 0$	$\beta = 0.1$	$\beta = 1$	$\beta = 2$	$\beta = 10$
Optimality	0.2365	0.2094	0.0701	0.0438	0.0164
Feasibility	99.90%	99.88%	99.78%	99.66%	99.61%

- As K increases, the sample converges to the constraint.
- As β increases, the target $\pi(\boldsymbol{x})$ concentrates around $\boldsymbol{x}^*$.

Quantitative Evaluation (FFHQ)

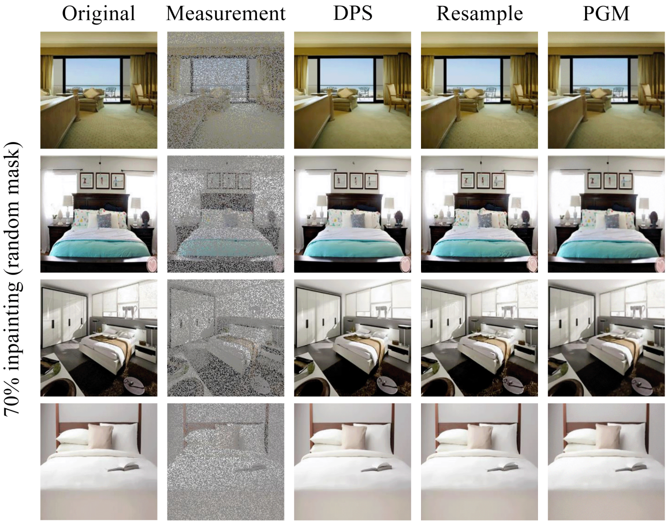
FFHQ	Super resolution 4×			Inpainting (random 70%)			Gaussian deblurring		
Method	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
DPS	28.47	0.793	0.175	32.32	0.897	0.106	27.70	0.774	0.169
MCG	23.74	0.673	0.223	24.89	0.731	0.178	25.33	0.668	0.371
ADMM-PnP	21.30	0.760	0.303	15.87	0.608	0.308	21.23	0.675	0.399
DDRM	27.51	0.753	0.257	24.97	0.680	0.287	26.51	0.702	0.299
DMPS	27.21	0.766	0.181	28.17	0.814	0.150	26.04	0.699	0.227
Latent-DPS	24.65	0.609	0.344	27.08	0.727	0.270	25.98	0.704	0.258
PSLD-LDM	27.22	0.705	0.267	25.61	0.630	0.270	20.08	0.400	0.422
ReSample	28.90	0.804	0.164	31.34	0.890	0.099	28.73	0.794	0.201
PGM (Our)	**29.82**	**0.956**	**0.158**	**32.48**	**0.962**	**0.092**	**30.67**	**0.959**	**0.148**

- PGM achieves a superior trade-off between numerical fidelity (PSNR) and perceptual similarity (SSIM and LPIPS).

Quantitative Evaluation (CelebA-HQ)

CelebA-HQ	Super resolution 4×			Inpainting (random 70%)			Gaussian deblurring		
Method	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
DPS	28.41	0.782	0.173	32.48	0.899	0.102	28.36	0.772	0.175
MCG	25.92	0.740	0.193	29.53	0.847	0.134	15.85	0.536	0.517
ADMM-PnP	21.08	0.631	0.304	15.40	0.342	0.627	20.98	0.602	0.289
DDRM	29.49	0.817	0.151	27.69	0.798	0.166	26.88	0.747	0.193
DMPS	28.48	0.811	0.147	28.84	0.826	0.175	26.45	0.726	0.206
Latent-DPS	26.83	0.690	0.272	26.23	0.703	0.226	27.42	0.729	0.205
PSLD-LDM	27.61	0.704	0.209	27.07	0.689	0.260	24.21	0.548	0.323
ReSample	30.45	0.832	0.144	32.77	0.903	0.082	30.69	0.832	0.148
PGM (Our)	**30.84**	**0.951**	**0.131**	**33.41**	**0.960**	**0.076**	**31.52**	**0.957**	**0.129**

Visual Results (LSUN-Bedroom)

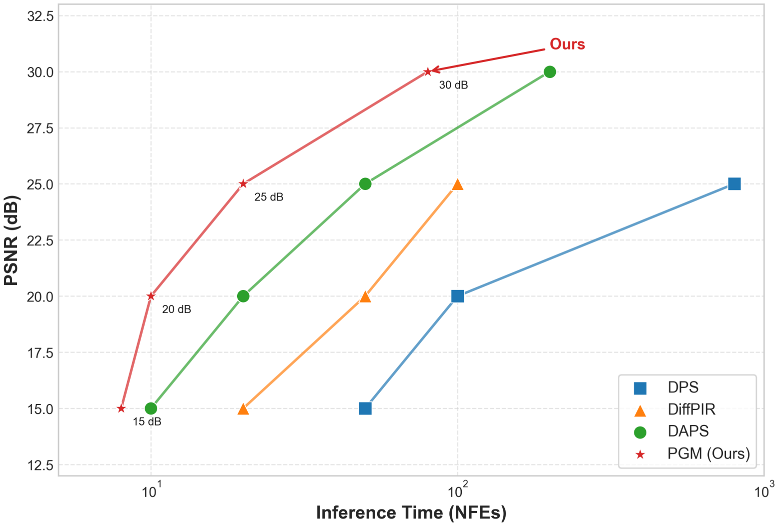


Memory and Sampling Time

Method	Model size	Total memory	Sampling time
DPS	1953 MB	5369 MB	36 s
DMPS	1953 MB	7168 MB	43 s
PSLD	3969 MB	9485 MB	65 s
ReSample	3969 MB	5009 MB	54 s
PGM (Ours)	**2055 MB**	**4682 MB**	**4 s**

- More memory efficiency by **avoiding backpropagation**.
- Dramatic $9\times$ **speedup** in the sampling time.

Trade-off between Quality and Efficiency



Quantitative Evaluation of Nonlinear Deblurring.

Nonlinear Deblurring		FFHQ			ImageNet		
Methods	Type	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
PGM(Ours)	Pixel	28.47	0.756	0.231	26.39	0.727	0.280
DPS		23.39	0.623	0.278	22.49	0.591	0.306
RED-diff		**30.86**	**0.795**	0.160	**30.07**	**0.754**	0.211
DAPS		28.29	0.783	**0.155**	27.73	0.724	**0.169**
ReSample	Latent	28.24	0.742	0.185	26.20	0.653	0.206
LatentDAPS		28.11	0.713	0.235	25.34	0.615	0.314

- Performance degrades for nonlinear inverse problem.
- Error: first-order approximation in operator splitting:

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^\top (\boldsymbol{x} - \boldsymbol{x}_k).$$

Summary

- PGM, a proximal-based generative modeling framework that addresses fundamental challenges in inverse problems.
- PGM enables a new pre-trained proximal diffusion sampling mechanism driven by the Moreau score.
- Non-asymptotic convergence under convexity and SOTA empirical performance across multiple tasks.
- PGM not only advances the theoretical understanding of generative modeling but also offers a practical, efficient, and highly effective tool for inverse problems.

Future Work

- Further enhancement for nonconvex priors and nonlinear inverse problems.
- More efficient network architectures for proximal operator.
- Theoretical analysis under weaker assumptions.

THANKS FOR LISTENING!

Paper

Code