Information-Theoretic Diffusion

Xianghao Kong[1*], Rob Brekelmans[2], Greg Ver Steeg[1]

[1]University of California, Riverside

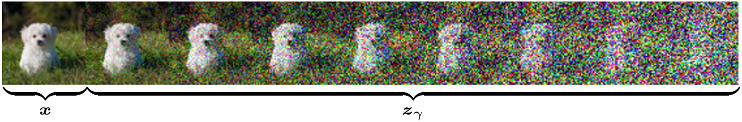
[2]Vector Institute

[*]Presenter

ICLR 2023

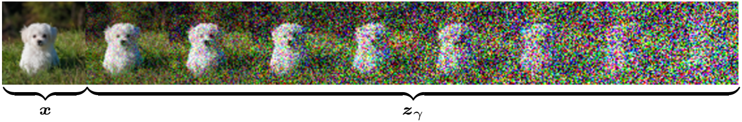
I-MMSE: Relating Information and Estimation

- Gaussian noise channel: $\boldsymbol{z}_\gamma = \sqrt{\gamma}\boldsymbol{x} + \boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbb{I})$, where γ is Signal-to-Noise Ratio (SNR).



I-MMSE: Relating Information and Estimation

- Gaussian noise channel: $\boldsymbol{z}_\gamma = \sqrt{\gamma}\boldsymbol{x} + \boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbb{I})$, where γ is Signal-to-Noise Ratio (SNR).



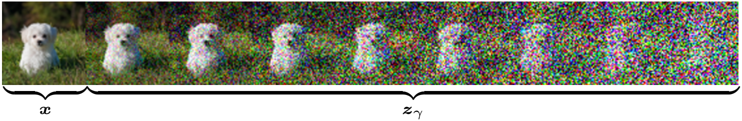
- Minimum Mean Square Error (MMSE):

$$\text{mmse}(\gamma) \equiv \min_{\hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, \gamma)} \mathbb{E}_{p(\boldsymbol{z}_\gamma, \boldsymbol{x})}\left[\|\boldsymbol{x} - \hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, \gamma)\|_2^2\right],$$

and its optimal denoiser is $\hat{\boldsymbol{x}}^*(\boldsymbol{z}_\gamma, \gamma) \equiv \arg \text{mmse}(\gamma)$.

I-MMSE: Relating Information and Estimation

- Gaussian noise channel: $\boldsymbol{z}_\gamma = \sqrt{\gamma}\boldsymbol{x} + \boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbb{I})$, where γ is Signal-to-Noise Ratio (SNR).



- Minimum Mean Square Error (MMSE):

$$\text{mmse}(\gamma) \equiv \min_{\hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, \gamma)} \mathbb{E}_{p(\boldsymbol{z}_\gamma, \boldsymbol{x})}\left[\|\boldsymbol{x} - \hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, \gamma)\|_2^2\right],$$

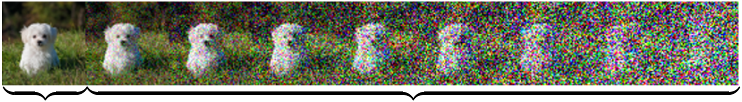
  and its optimal denoiser is $\hat{\boldsymbol{x}}^*(\boldsymbol{z}_\gamma, \gamma) \equiv \arg \text{mmse}(\gamma)$.

- Information and MMSE (I-MMSE) relations (Guo et al., 2005):

$$\frac{d}{d\gamma} \underbrace{I(\boldsymbol{x}; \boldsymbol{z}_\gamma)}_{\substack{\text{mutual} \\ \text{information}}} = \frac{1}{2}\text{mmse}(\gamma)$$

I-MMSE: Relating Information and Estimation

- Gaussian noise channel: $\boldsymbol{z}_\gamma = \sqrt{\gamma}\boldsymbol{x} + \boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbb{I})$, where γ is Signal-to-Noise Ratio (SNR).



- Minimum Mean Square Error (MMSE):

$$\mathrm{mmse}(\gamma) \equiv \min_{\hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, \gamma)} \mathbb{E}_{p(\boldsymbol{z}_\gamma, \boldsymbol{x})}\left[\|\boldsymbol{x} - \hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, \gamma)\|_2^2\right],$$

  and its optimal denoiser is $\hat{\boldsymbol{x}}^*(\boldsymbol{z}_\gamma, \gamma) \equiv \arg \mathrm{mmse}(\gamma)$.

- Information and MMSE (I-MMSE) relations (Guo et al., 2005):

$$\frac{d}{d\gamma} I(\boldsymbol{x}; \boldsymbol{z}_\gamma) = \frac{1}{2}\mathrm{mmse}(\gamma)$$

- Pointwise I-MMSE relations:

$$\frac{d}{d\gamma} D_{KL}[p(\boldsymbol{z}_\gamma|\boldsymbol{x}) \parallel p(\boldsymbol{z}_\gamma)] = \frac{1}{2}\,\mathrm{mmse}(\boldsymbol{x}, \gamma)$$

  where $\mathrm{mmse}(\boldsymbol{x}, \gamma) \equiv \mathbb{E}_{p(\boldsymbol{z}_\gamma|\boldsymbol{x})}\left[\|\boldsymbol{x} - \hat{\boldsymbol{x}}^*(\boldsymbol{z}_\gamma, \gamma)\|_2^2\right]$.

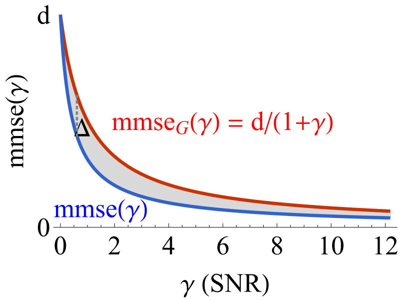
Probability = Optimal Denoising

- Thermodynamic Integration: $\int_{\gamma_0}^{\gamma_1} d\gamma \frac{d}{d\gamma} f(\gamma) = f(\gamma_1) - f(\gamma_0)$.
- Define $f(\boldsymbol{x}, \gamma) \equiv \underbrace{D_{KL}[p(\boldsymbol{z}_\gamma|\boldsymbol{x}) \parallel p_G(\boldsymbol{z}_\gamma)]}_{\text{KL for Gaussian distribution}} - \underbrace{D_{KL}[p(\boldsymbol{z}_\gamma|\boldsymbol{x}) \parallel p(\boldsymbol{z}_\gamma)]}_{\text{KL for data distribution}}$.

Probability = Optimal Denoising

- Thermodynamic Integration: $\int_{\gamma_0}^{\gamma_1} d\gamma \frac{d}{d\gamma} f(\gamma) = f(\gamma_1) - f(\gamma_0)$.
- Define $f(\boldsymbol{x}, \gamma) \equiv \underbrace{D_{KL}[p(\boldsymbol{z}_\gamma|\boldsymbol{x}) \parallel p_G(\boldsymbol{z}_\gamma)]}_{\text{KL for Gaussian distribution}} - \underbrace{D_{KL}[p(\boldsymbol{z}_\gamma|\boldsymbol{x}) \parallel p(\boldsymbol{z}_\gamma)]}_{\text{KL for data distribution}}$.

$$-\log p(\boldsymbol{x}) = \underbrace{\overbrace{\frac{d}{2}\log(2\pi e)}^{\textbf{constant!}}}_{\text{Gaussian entropy}} - \frac{1}{2}\int_0^\infty d\gamma \left(\underbrace{\overbrace{\frac{d}{1+\gamma}}^{\textbf{constant!}}}_{\text{MMSE for Gaussian}} - \underbrace{\text{mmse}(\boldsymbol{x}, \gamma)}_{\text{MMSE for data}} \right)$$



Diffusion Math Comparison

- Variational bound:

$$-\log p(\boldsymbol{x}) \quad \leq \quad \underbrace{L_0}_{\substack{\text{discrete}\\\text{reconstruction}}} \quad + \underbrace{L_{1:T-1}}_{\text{MSE terms}} + \underbrace{L_T}_{\text{prior}}$$

  - Discrete sum only approximated integral
  - Superfluous, non-MSE terms in variational bound

Diffusion Math Comparison

- Variational bound:

$$-\log p(\boldsymbol{x}) \quad {\color{red}\leq} \quad \underbrace{L_0}_{\substack{\text{discrete}\\\text{reconstruction}}} \quad + \underbrace{L_{1:T-1}}_{\text{MSE terms}} + \underbrace{L_T}_{\text{prior}}$$

  - Discrete sum only approximated integral
  - Superfluous, non-MSE terms in variational bound

Diffusion Math Comparison

- Variational bound:

$$-\log p(\boldsymbol{x}) \leq \underbrace{L_0}_{\substack{\text{discrete}\\ \text{reconstruction}}} + \underbrace{L_{1:T-1}}_{\text{MSE terms}} + \underbrace{L_T}_{\text{prior}}$$

  - Discrete sum only approximated integral
  - Superfluous, non-MSE terms in variational bound

- Information-theoretic bound (ours):

$$-\log p(\boldsymbol{x}) = \text{constant} + \frac{1}{2}\int_0^\infty \text{mmse}(\boldsymbol{x}, \gamma) d\gamma$$

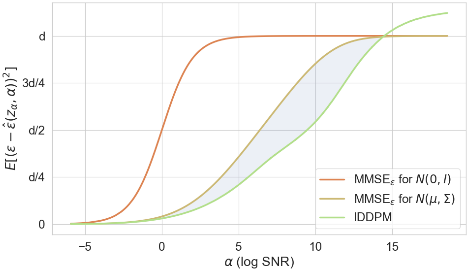
  - Analytic solution
  - MMSE gap reveals relevant SNR ranges

Density Estimation

- CIFAR-10 test data
- Using a SOTA U-Net from "Improved Denoising Diffusion Probabilistic Models" paper.

Table: $\mathbb{E}\left[-\log p(\boldsymbol{x})\right]$ (bits/dimension)

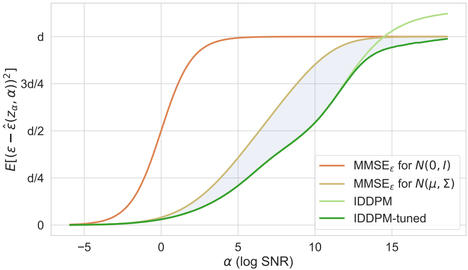
		Test-time estimate	
Model	Training Objective	Variational Bound	IT Bound (ours)
IDDPM	Variational	-4.05	**-4.09**



- CIFAR-10 test data
- Using a SOTA U-Net from "Improved Denoising Diffusion Probabilistic Models" paper.

Table: $\mathbb{E}\left[-\log p(\boldsymbol{x})\right]$ (bits/dimension)

		Test-time estimate	
Model	Training Objective	Variational Bound	IT Bound (ours)
IDDPM	Variational	-4.05	**-4.09**
IDDPM(tune)	Info-Theoretic	-3.85	**-4.28**



Contributions

- A new I-MMSE relation and an exact diffusion bound:

$$-\log p(\boldsymbol{x}) = \frac{1}{2}\int_0^\infty \text{mmse}(\boldsymbol{x},\gamma)d\gamma + \text{constant terms.}$$

- The same optimization problem can also be exactly related to *discrete* probability mass.

$$-\log P(\boldsymbol{x}) = \frac{1}{2}\int_0^\infty \text{mmse}(\boldsymbol{x},\gamma)d\gamma$$

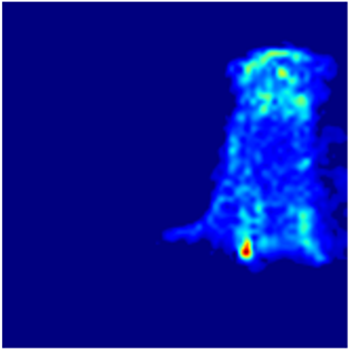
- Our approach allows us to fine-tune and *ensemble* existing diffusion models to achieve better NLLs.

- Code to reproduce experiments is provided at
`https://github.com/kxh001/ITdiffusion`

- Simplified demonstration code is at
`https://github.com/gregversteeg/InfoDiffusionSimple`

Future Work - Information Decomposition

Query image



Informative pixels about “dog” extracted from Stable Diffusion model.