# Estimating the Optimal Covariance with Imperfect Mean in Diffusion Probabilistic Models

https://github.com/baofff/Extended-Analytic-DPM

Tsinghua University

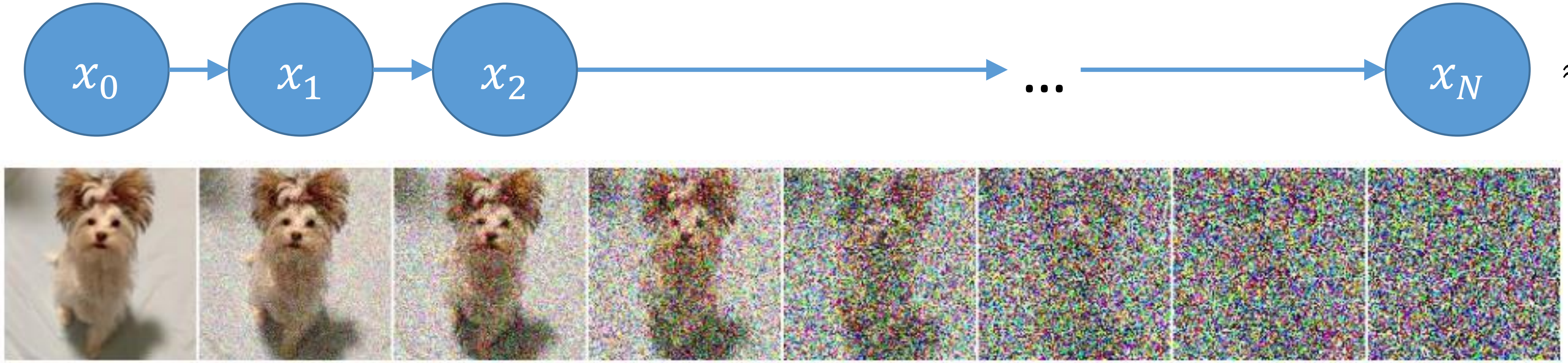
Fan Bao, Chongxuan Li, Jiacheng Sun, Jun Zhu, Bo Zhang

# Diffusion Probabilistic Models (DPMs)

*Ho et al. Denoising diffusion probabilistic models (DDPM), Neurips 2020.*
*Song et al. Score-based generative modeling through stochastic differential equations, ICLR 2021.*

- Diffusion process gradually injects noise to data
- Described by a Markov chain: $q(x_0, \dots, x_N) = q(x_0)q(x_1|x_0) \dots q(x_N|x_{N-1})$

Transition of diffusion: $q(x_n|x_{n-1}) = N(\sqrt{\alpha_n}x_{n-1}, \beta_n I)$ $\quad \alpha_n = 1 - \beta_n$



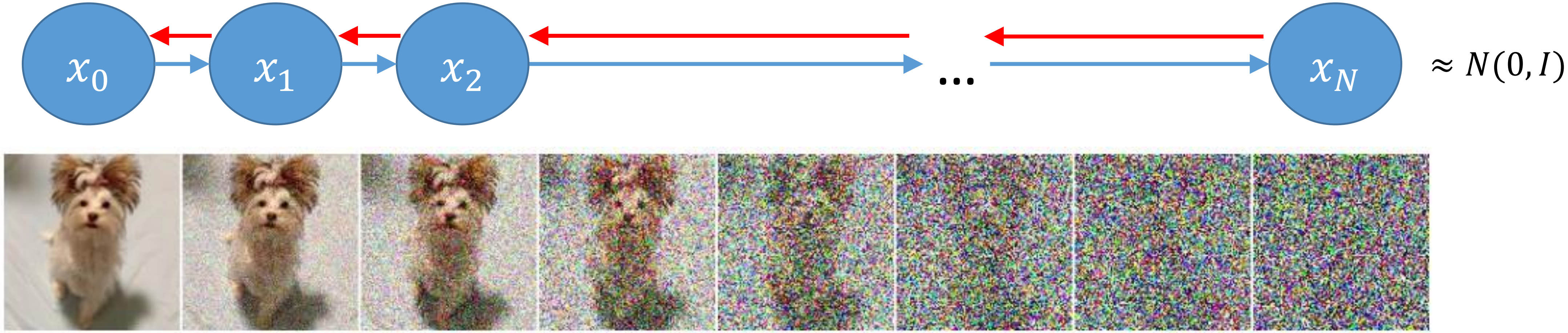

$\approx N(0, I)$

Diffusion process: $q(x_0, \dots, x_N) = q(x_0)q(x_1|x_0) \dots q(x_N|x_{N-1})$

Demo Images from *Song et al. Score-based generative modeling through stochastic differential equations, ICLR 2021.*

- Diffusion process in the reverse direction ⇔ denoising process
- Reverse factorization: $q(x_0, \dots, x_N) = q(x_0|x_1) \dots q(x_{N-1}|x_N)q(x_N)$

Transition of denoising: $q(x_{n-1}|x_n) = ?$




Diffusion process: $q(x_0, \dots, x_N) = q(x_0)q(x_1|x_0) \dots q(x_N|x_{N-1})$

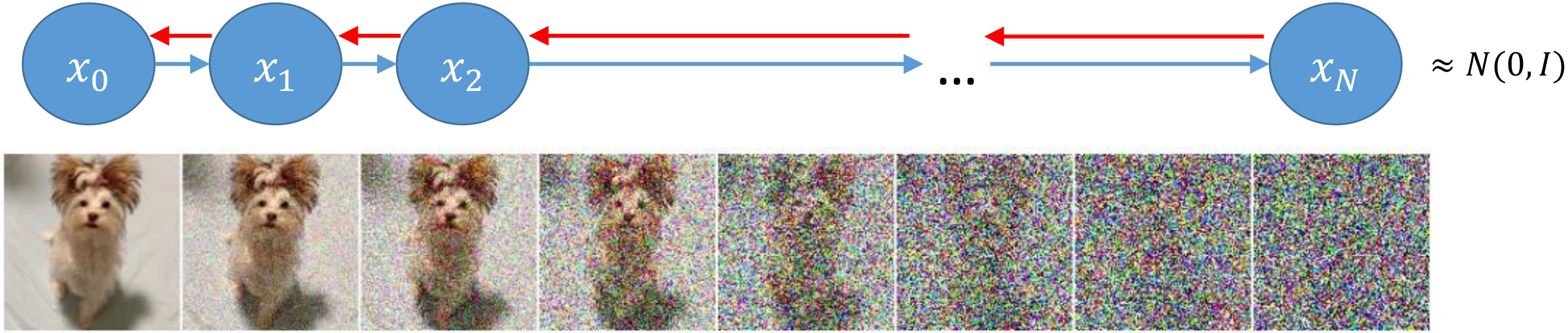
$= q(x_0|x_1) \dots q(x_{N-1}|x_N)q(x_N)$

- Approximate diffusion process in the reverse direction

Model transition: $p(x_{n-1}|x_n) = N(\mu_n(x_n), \Sigma_n(x_n))$

↓ approximate

Transition of denoising: $q(x_{n-1}|x_n) = ?$




Diffusion process: $q(x_0, \ldots, x_N) = q(x_0)q(x_1|x_0) \ldots q(x_N|x_{N-1})$

$= q(x_0|x_1) \ldots q(x_{N-1}|x_N)q(x_N)$

The model: $p(x_0, \ldots, x_N) = p(x_0|x_1) \ldots p(x_{N-1}|x_N)p(x_N)$

- We hope $q(x_0, \dots, x_N) \approx p(x_0, \dots, x_N)$ $\quad p(x_{n-1}|x_n) = N(\mu_n(x_n), \Sigma_n(x_n))$
- Achieved by minimizing their KL divergence (i.e., maximizing the ELBO)

min KL $\qquad$ max ELBO

$$\min_{\mu_n(\cdot),\Sigma_n(\cdot)} KL(q(x_{0:N})||p(x_{0:N})) \Leftrightarrow \max_{\mu_n(\cdot),\Sigma_n(\cdot)} \mathrm{E}_q \log \frac{p(x_{0:N})}{q(x_{1:N}|x_0)}$$

(**Analytic-DPM**) Suppose $\Sigma_n(x_n) = \sigma_n^2$. The optimal solution is

$$\mu_n^*(x_n) = \frac{1}{\sqrt{\alpha_n}}\left(x_n - \frac{\beta_n}{\sqrt{\bar{\beta}_n}} \mathbf{E}_{q(x_0|x_n)}[\epsilon_n]\right),$$

$$\sigma_n^{*2} = \frac{\beta_n}{\alpha_n}\left(1 - \frac{\beta_n}{\bar{\beta}_n} \mathrm{E}_{q_n(x_n)} \frac{\left\|\mathbf{E}_{q(x_0|x_n)}[\epsilon_n]\right\|^2}{d}\right).$$

Noise prediction network:
$\hat{\epsilon}_n(x_n) \approx \mathbf{E}_{q(x_0|x_n)}[\epsilon_n]$

- ➢ Isotropic variance is simple
- ➢ Irrelevant to state $x_n$
- ➢ Assume the mean is optimal

*Bao et al. Analytic-DPM: an Analytic Estimate of the Optimal Reverse Variance in Diffusion Probabilistic Models, ICLR 2022.*

# Estimating the Optimal Covariance with Imperfect Mean in Diffusion Probabilistic Models

- What is the optimal diagonal covariance $\Sigma_n(x_n) = \mathrm{diag}(\sigma_n^2(x_n))$ ?

**Theorem 1.** Suppose $\Sigma_n(x_n) = \mathrm{diag}(\sigma_n^2(x_n))$. The optimal solution is

$$\mu_n^*(x_n) = \frac{1}{\sqrt{\alpha_n}}\left(x_n - \frac{\beta_n}{\sqrt{\overline{\beta}_n}}\mathbf{E}_{q(x_0|x_n)}[\epsilon_n]\right),$$

$$\sigma_n^*(x_n)^2 = \frac{\overline{\beta}_{n-1}}{\overline{\beta}_n}\beta_n + \frac{\beta_n^2}{\overline{\beta}_n\alpha_n}\Big(\underbrace{\mathbf{E}_{q(x_0|x_n)}[\epsilon_n^2]}_{\approx h_n(x_n)} - \underbrace{\mathbf{E}_{q(x_0|x_n)}[\epsilon_n]^2}_{\approx \hat{\epsilon}_n(x_n)^2}\Big).$$

predict SN: $\min_{h_n} \mathbf{E}_{q(x_0,x_n)} \| h_n(x_n) - \epsilon_n^2 \|^2$

squared noise (SN)

See a more general version of Theorem 1 for more general $q(x_{0:N})$ in the full paper
See extension to score-based SDE (Song et al.) in the full paper

- In practice, $\hat{\epsilon}_n(x_n)$ and $\mu_n(x_n)$ are not optimal
- What is the optimal diagonal covariance $\Sigma_n(x_n) = \mathrm{diag}(\sigma_n^2(x_n))$ ?

**Theorem 2.** Suppose $\Sigma_n(x_n) = \mathrm{diag}(\sigma_n^2(x_n))$. For any mean $\mu_n(x_n)$ parameterized by $\hat{\epsilon}_n(x_n)$, the optimal covariance is

$$\tilde{\sigma}_n^*(x_n)^2 = \frac{\overline{\beta}_{n-1}}{\overline{\beta}_n}\beta_n + \frac{\beta_n^2}{\overline{\beta}_n\alpha_n}\underbrace{\mathbf{E}_{q(x_0|x_n)}\left[\left(\epsilon_n - \hat{\epsilon}_n(x_n)\right)^2\right]}_{\approx\, g_n(x_n)}.$$

predict NPR: $\min_{g_n}\mathbf{E}_{q(x_0,x_n)}\left\|h_n(x_n) - \boxed{\left(\epsilon_n - \hat{\epsilon}_n(x_n)\right)^2}\right\|^2$

noise prediction residual (NPR): $\left(\epsilon_n - \hat{\epsilon}_n(x_n)\right)^2$

# Experimental Results

- Likelihood results
- NPR-DPM (predicting NPR) consistently outperforms Analytic-DPM

| | | CIFAR10 (LS) | | | | | | CIFAR10 (CS) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # TIMESTEPS $K$ | | 10 | 25 | 50 | 100 | 200 | 1000 | 10 | 25 | 50 | 100 | 200 | 1000 |
| ET | DDPM, $\tilde{\beta}_n$ | 74.95 | 24.98 | 12.01 | 7.08 | 5.03 | 3.73 | 75.96 | 24.94 | 11.96 | 7.04 | 4.95 | 3.60 |
| | DDPM, $\beta_n$ | 6.99 | 6.11 | 5.44 | 4.86 | 4.39 | 3.75 | 6.51 | 5.55 | 4.92 | 4.41 | 4.03 | 3.54 |
| | A-DDPM | 5.47 | 4.79 | 4.38 | 4.07 | 3.84 | 3.59 | 5.08 | 4.45 | 4.09 | 3.83 | 3.64 | 3.42 |
| | NPR-DDPM | **5.40** | **4.64** | **4.25** | **3.98** | **3.79** | **3.57** | **5.03** | **4.33** | **3.99** | **3.76** | **3.59** | **3.41** |
| OT | DDPM, $\beta_n$ | 5.38 | 4.34 | 3.97 | 3.82 | 3.77 | 3.75 | 5.51 | 4.30 | 3.86 | 3.65 | 3.57 | 3.54 |
| | A-DDPM | 4.11 | 3.68 | 3.61 | 3.59 | 3.59 | 3.59 | 3.99 | 3.56 | 3.47 | 3.44 | 3.43 | 3.42 |
| | NPR-DDPM | **3.91** | **3.64** | **3.59** | **3.58** | **3.57** | **3.57** | **3.88** | **3.52** | **3.45** | **3.42** | **3.41** | **3.41** |

| | | CELEBA 64X64 | | | | | | IMAGENET 64X64 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # TIMESTEPS $K$ | | 10 | 25 | 50 | 100 | 200 | 1000 | 25 | 50 | 100 | 200 | 400 | 4000 |
| ET | DDPM, $\tilde{\beta}_n$ | 33.42 | 13.09 | 7.14 | 4.60 | 3.45 | 2.71 | 105.87 | 46.25 | 22.02 | 12.10 | 7.59 | 3.89 |
| | DDPM, $\beta_n$ | 6.67 | 5.72 | 4.98 | 4.31 | 3.74 | 2.93 | 5.81 | 5.20 | 4.70 | 4.31 | 4.04 | 3.65 |
| | A-DDPM | 4.54 | 3.89 | 3.48 | 3.16 | 2.92 | 2.66 | 4.78 | 4.42 | 4.15 | 3.95 | 3.81 | 3.61 |
| | NPR-DDPM | **4.46** | **3.78** | **3.40** | **3.11** | **2.89** | **2.65** | **4.66** | **4.22** | **3.96** | **3.80** | **3.71** | **3.60** |
| OT | DDPM, $\beta_n$ | 4.76 | 3.58 | 3.16 | 2.99 | 2.94 | 2.93 | 4.56 | 4.09 | 3.84 | 3.73 | 3.68 | 3.65 |
| | A-DDPM | 2.97 | 2.71 | 2.67 | 2.66 | 2.66 | 2.66 | 3.83 | 3.70 | 3.64 | 3.62 | 3.62 | 3.61 |
| | NPR-DDPM | **2.88** | **2.69** | **2.66** | **2.66** | **2.65** | **2.65** | **3.73** | **3.65** | **3.62** | **3.60** | **3.60** | **3.60** |

- Sample quality results (FID)
- Both NPR-DPM & SN-DPM outperform Analytic-DPM

| | CIFAR10 (LS) | | | | | | CIFAR10 (CS) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # TIMESTEPS $K$ | 10 | 25 | 50 | 100 | 200 | 1000 | 10 | 25 | 50 | 100 | 200 | 1000 |
| DDPM, $\tilde{\beta}_n$ | 44.45 | 21.83 | 15.21 | 10.94 | 8.23 | 5.11 | 34.76 | 16.18 | 11.11 | 8.38 | 6.66 | 4.92 |
| DDPM, $\beta_n$ | 233.41 | 125.05 | 66.28 | 31.36 | 12.96 | **3.04** | 205.31 | 84.71 | 37.35 | 14.81 | 5.74 | **3.34** |
| A-DDPM | 34.26 | 11.60 | 7.25 | 5.40 | 4.01 | 4.03 | 22.94 | 8.50 | 5.50 | 4.45 | 4.04 | 4.31 |
| NPR-DDPM | 32.35 | 10.55 | 6.18 | 4.52 | 3.57 | 4.10 | 19.94 | 7.99 | 5.31 | 4.52 | 4.10 | 4.27 |
| SN-DDPM | **24.06** | **6.91** | **4.63** | **3.67** | **3.31** | 3.65 | **16.33** | **6.05** | **4.17** | **3.83** | **3.72** | 4.07 |
| DDIM | 21.31 | 10.70 | 7.74 | 6.08 | 5.07 | 4.13 | 34.34 | 16.68 | 10.48 | 7.94 | 6.69 | 4.89 |
| A-DDIM | 14.00 | 5.81 | 4.04 | 3.55 | 3.39 | 3.74 | 26.43 | 9.96 | 6.02 | 4.88 | 4.92 | 4.66 |
| NPR-DDIM | 13.34 | 5.38 | 3.95 | 3.53 | 3.42 | 3.72 | 22.81 | 9.47 | 6.04 | 5.02 | 5.06 | 4.62 |
| SN-DDIM | **12.19** | **4.28** | **3.39** | **3.23** | **3.22** | **3.65** | **17.90** | **7.36** | **5.16** | **4.63** | **4.63** | **4.51** |

| | CELEBA 64x64 | | | | | | IMAGENET 64x64 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # TIMESTEPS $K$ | 10 | 25 | 50 | 100 | 200 | 1000 | 25 | 50 | 100 | 200 | 400 | 4000 |
| DDPM, $\tilde{\beta}_n$ | 36.69 | 24.46 | 18.96 | 14.31 | 10.48 | 5.95 | 29.21 | 21.71 | 19.12 | 17.81 | 17.48 | 16.55 |
| DDPM, $\beta_n$ | 294.79 | 115.69 | 53.39 | 25.65 | 9.72 | **3.16** | 170.28 | 83.86 | 45.04 | 28.39 | 21.38 | 16.38 |
| A-DDPM | 28.99 | 16.01 | 11.23 | 8.08 | 6.51 | 5.21 | 32.56 | 22.45 | 18.80 | 17.16 | 16.40 | 16.34 |
| NPR-DDPM | 28.37 | 15.74 | 10.89 | 8.23 | 7.03 | 5.33 | 28.27 | 20.89 | 18.06 | 16.96 | **16.32** | 16.38 |
| SN-DDPM | **20.60** | **12.00** | **7.88** | **5.89** | **5.02** | 4.42 | **27.58** | **20.74** | **18.04** | **16.61** | 16.37 | **16.22** |
| DDIM | 20.54 | 13.45 | 9.33 | 6.60 | 4.96 | 3.40 | 26.06 | 20.10 | 18.09 | 17.84 | 17.74 | 19.00 |
| A-DDIM | 15.62 | 9.22 | 6.13 | 4.29 | 3.46 | 3.13 | **25.98** | **19.23** | 17.73 | 17.49 | 17.44 | 18.98 |
| NPR-DDIM | 14.98 | 8.93 | 6.04 | 4.27 | 3.59 | 3.15 | 28.84 | 19.62 | 17.63 | 17.42 | 17.30 | 18.91 |
| SN-DDIM | **10.20** | **5.48** | **3.83** | **3.04** | **2.85** | **2.90** | 28.07 | 19.38 | **17.53** | **17.23** | **17.23** | **18.89** |

- Sample quality results (FID)
- Both NPR-DPM & SN-DPM outperform Analytic-DPM

| | CIFAR10 (VP SDE) | | | | | |
|---|---|---|---|---|---|---|
| # TIMESTEPS $K$ | 10 | 25 | 50 | 100 | 200 | 1000 |
| EULER-MARUYAMA | 292.20 | 170.17 | 90.79 | 47.46 | 21.92 | **2.55** |
| ANCESTRAL SAMPLING | 235.28 | 129.29 | 68.52 | 31.99 | 12.81 | 2.72 |
| PROBABILITY FLOW | 107.74 | 21.34 | 7.78 | 4.33 | 3.27 | 2.82 |
| A-DPM | 35.10 | 11.57 | 6.54 | 4.71 | 3.61 | 2.98 |
| NPR-DPM | 33.70 | 10.44 | 5.83 | 3.97 | 3.05 | 3.04 |
| SN-DPM | **25.30** | **7.34** | **4.46** | **3.27** | **2.83** | 2.71 |

# Thanks!