PriorGrad: Improving Conditional Denoising Diffusion Models with Data-Dependent Adaptive Prior

Sang-gil Lee, Heeseung Kim, Chaehun Shin, Xu Tan, Chang Liu, Qi Meng, Tao Qin, Wei Chen, Sungroh Yoon, Tie-Yan Liu

Data Science & Artificial Intelligence Laboratory
Electrical and Computer Engineering
Seoul National University

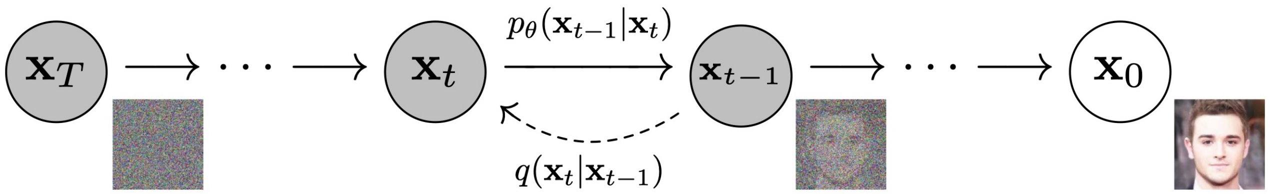
Diffusion model revolution… at what cost?

Diffusion model achieves SOTA results on generative tasks…

But how *efficient* are these models?

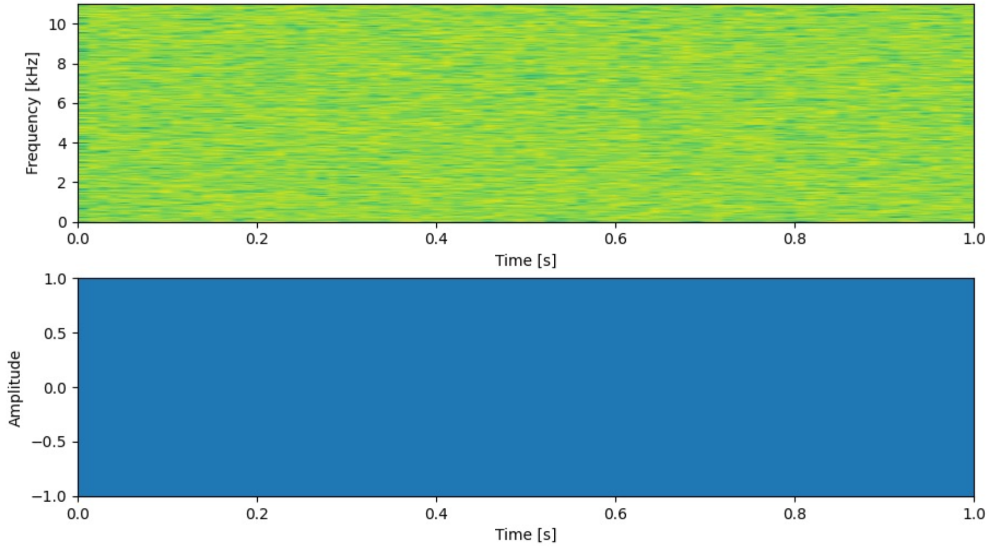
- Slow training brings a huge computational burden
- Slow sampling renders them not practical for real-world application

Why?: standard Gaussian prior is not efficient enough!



Source: Ho et al., 2020

Source: Dhariwal et al., 2021



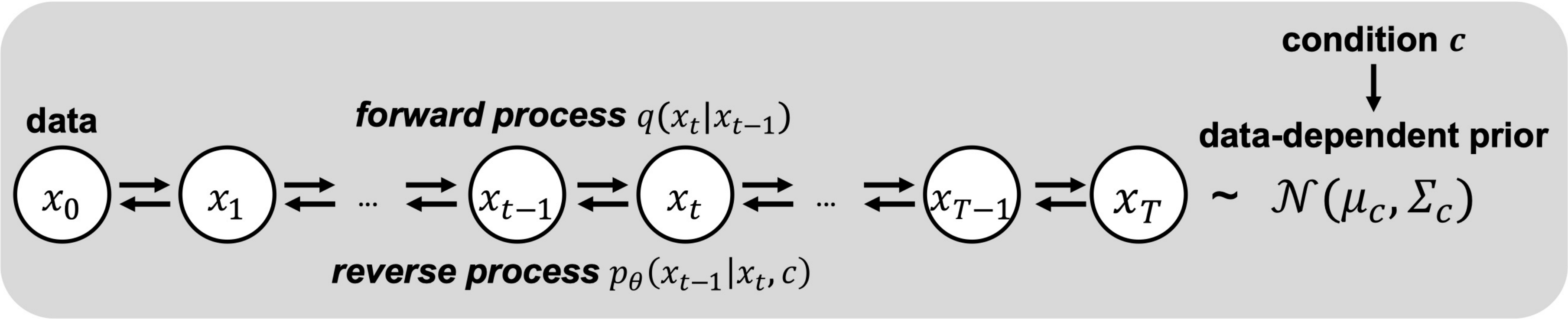
Source: https://github.com/mindslab-ai/wavegrad2

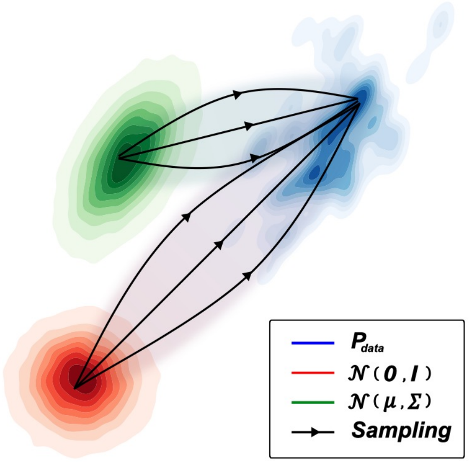
PriorGrad investigates the following question

For a conditional diffusion-based model, can we formulate a more informative prior without incorporating additional computational or parameter complexity?

PriorGrad presents efficiency for conditional diffusion models for free!

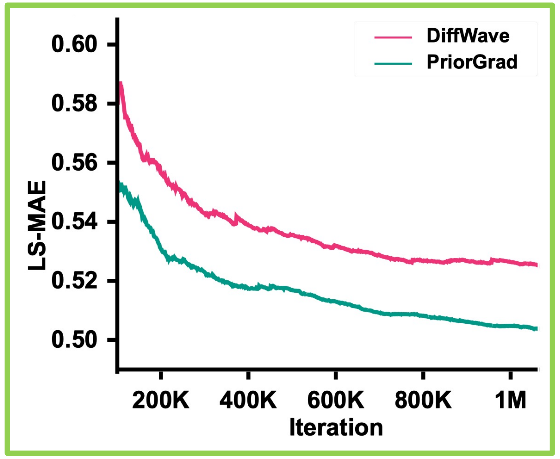
- Introduce a data-dependent prior based on the conditional information
- Construct a conditional non-standard Gaussian prior for training & inference
- Faster training & sampling, and SOTA results on speech synthesis tasks





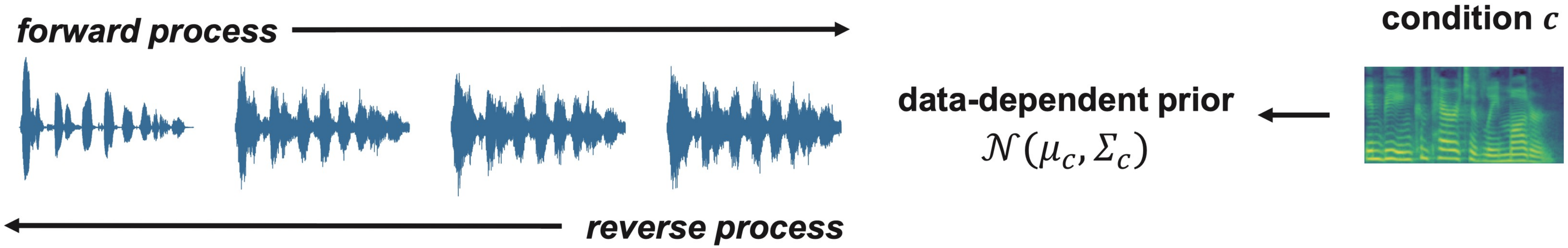
Faster training, sampling, and smaller capacity for diffusion models, for free!

PriorGrad enables faster training!



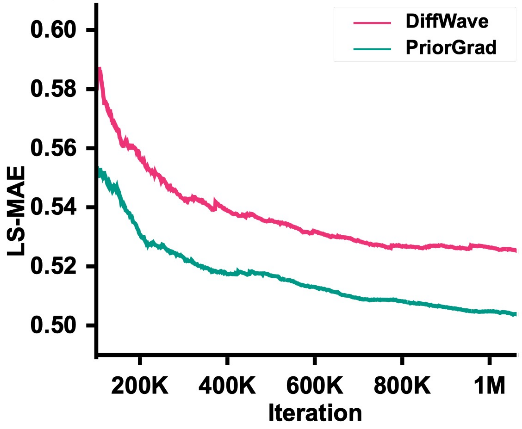
Method	T_{infer}	Training Steps 500K	Training Steps 1M
GT	-	4.42 ± 0.07	
DiffWave	6	3.98 ± 0.08	4.01 ± 0.08
	50	4.12 ± 0.08	4.12 ± 0.08
PriorGrad	6	4.02 ± 0.08	4.14 ± 0.08
	50	4.21 ± 0.08	**4.25** ± **0.08**

Method	Parameters Base (2.62M)	Parameters Small (1.23M)
GT	4.38 ± 0.08	
DiffWave	4.06 ± 0.08	3.90 ± 0.09
PriorGrad	**4.12** ± **0.08**	**4.02** ± **0.08**



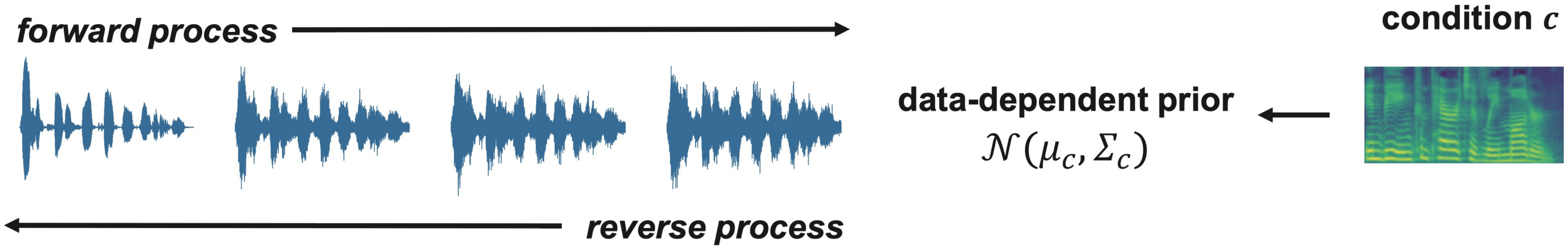
Faster training, sampling, and smaller capacity for diffusion models, for free!

PriorGrad enables faster sampling!



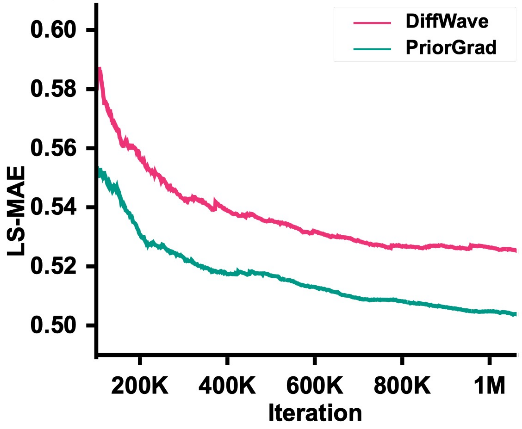
Method	T_{infer}	Training Steps 500K	1M
GT	-	4.42 ± 0.07	
DiffWave	6	3.98 ± 0.08	4.01 ± 0.08
	$\boxed{50}$	4.12 ± 0.08	$\boxed{4.12 \pm 0.08}$
PriorGrad	$\boxed{6}$	4.02 ± 0.08	$\boxed{4.14 \pm 0.08}$
	50	4.21 ± 0.08	$\mathbf{4.25} \pm \mathbf{0.08}$

Method	Parameters Base (2.62M)	Small (1.23M)
GT	4.38 ± 0.08	
DiffWave	4.06 ± 0.08	3.90 ± 0.09
PriorGrad	$\mathbf{4.12} \pm \mathbf{0.08}$	$\mathbf{4.02} \pm \mathbf{0.08}$



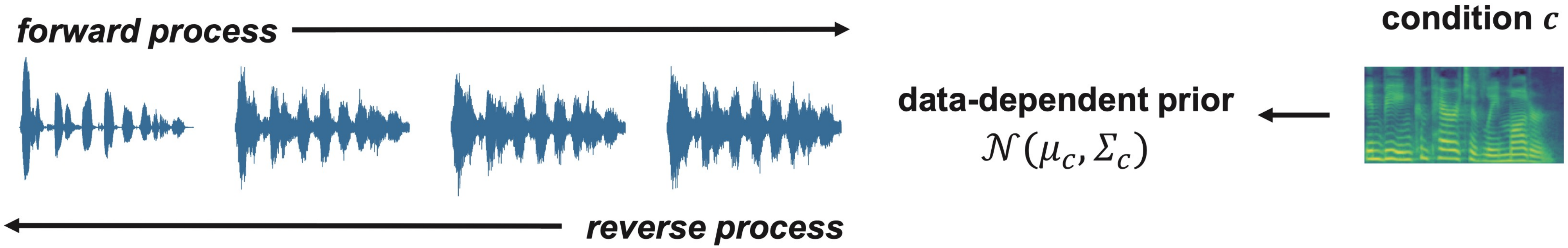
Faster training, sampling, and smaller capacity for diffusion models, for free!

PriorGrad enables smaller model capacity!



Method	T_{infer}	Training Steps 500K	Training Steps 1M
GT	-	4.42 ± 0.07	
DiffWave	6	3.98 ± 0.08	4.01 ± 0.08
	50	4.12 ± 0.08	4.12 ± 0.08
PriorGrad	6	4.02 ± 0.08	4.14 ± 0.08
	50	4.21 ± 0.08	**4.25 ± 0.08**

Method	Parameters Base (2.62M)	Parameters Small (1.23M)
GT	4.38 ± 0.08	
DiffWave	4.06 ± 0.08	3.90 ± 0.09
PriorGrad	**4.12 ± 0.08**	**4.02 ± 0.08**

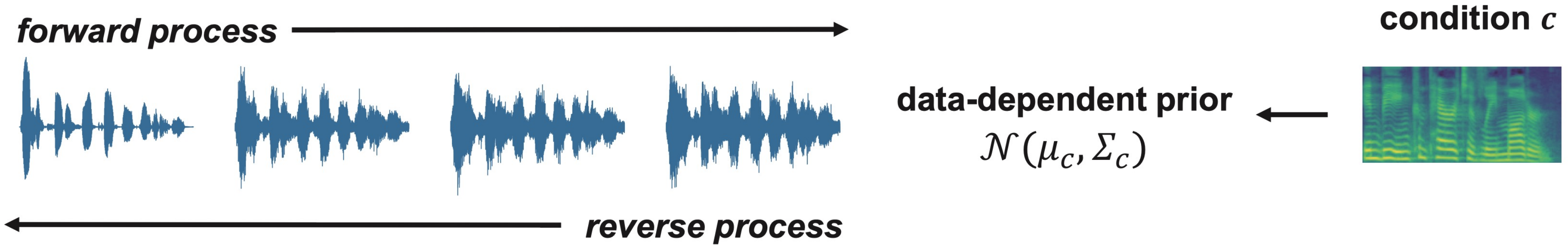


PriorGrad achieves the new SOTA likelihood-based vocoder

- Outperforms DiffWave (Kong *et al.*, 2021) and flow-based models (WaveGlow, WaveFlow)

Method	T_{infer}	MOS	RTF	Parameters
GT	-	4.60 ± 0.05	-	-
DiffWave / PriorGrad	6	4.10 ± 0.08 / 4.20 ± 0.08	0.1388	2.62M
	12	4.15 ± 0.08 / 4.29 ± 0.08	0.2780	
	50	4.19 ± 0.07 / $\mathbf{4.33 \pm 0.07}$	1.1520	
WaveGlow †	-	4.09 ± 0.08	0.0780	87.9M
WaveFlow †	-	4.01 ± 0.09	0.1759	22.3M
HiFi-GAN (V1) †	-	$\mathbf{4.44 \pm 0.05}$	**0.0068**	14.0M

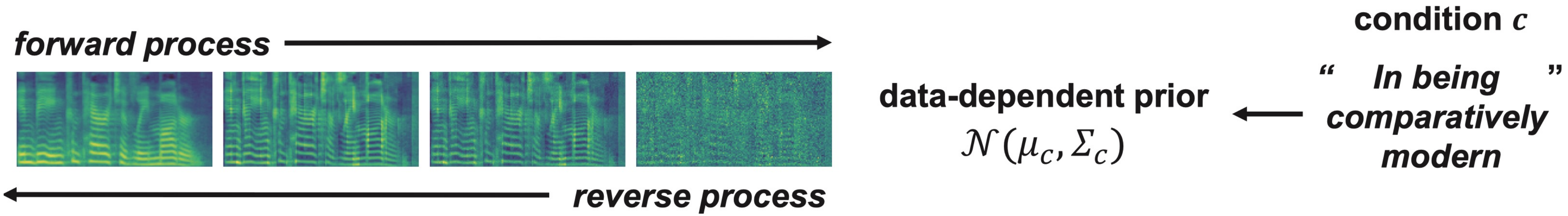
Method	LS-MAE (↓)	MR-STFT (↓)	MCD (↓)	F_0 RMSE (↓)	$S(x_T, x_0)$ (↓)	$S(\tilde{x}_0, x_0)$ (↓)
DiffWave	0.5264	1.0920	9.7822	16.4035	72698.62	1650.22
PriorGrad	**0.5048**	**0.9976**	**9.2820**	**15.5542**	**42236.93**	**1608.89**



PriorGrad achieves the new SOTA acoustic model

- Outperforms every well-known acoustic models (FastSpeech 2, Glow-TTS, Grad-TTS)
- High-quality sampling with just 2 inference steps

Method	T_{infer}	MOS	RTF	Parameters Encoder	Parameters Decoder
GT	-	4.65 ± 0.05	-	-	-
GT (Vocoder)	-	4.50 ± 0.06	-	-	-
Baseline / PriorGrad	2	2.80 ± 0.17 / 4.25 ± 0.08	**0.0069**	11.5M	3.5M
	6	3.67 ± 0.12 / 4.29 ± 0.07	0.0113		
	12	4.14 ± 0.08 / $\mathbf{4.39 \pm 0.08}$	0.0176		
Grad-TTS†	2	3.43 ± 0.15	0.0090	7.2M	7.6M
	10	$\mathbf{4.38 \pm 0.05}$	0.0308		
FastSpeech 2	-	4.19 ± 0.08	**0.0040**	11.5M	11.5M
Glow-TTS†	-	4.23 ± 0.08	0.0081	7.2M	21.4M



Theoretical benefits

PriorGrad can use a simpler model to represent the reverse diffusion

Proposition 2 *Let* $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{x}_0; \theta)$ *denote the* $-\mathrm{ELBO}$ *loss in Proposition 1. Suppose that* $\boldsymbol{\epsilon}_\theta$ *is a linear function. Under the constraint that* $\det(\boldsymbol{\Sigma}) = \det(\mathbf{I})$, *we have* $\min_\theta L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{x}_0; \theta) \leq \min_\theta L(\mathbf{0}, \mathbf{I}, \mathbf{x}_0; \theta)$.

PriorGrad guarantees faster convergence with the smaller condition number of the Hessian

$$\boxed{\frac{\lambda_{\max}(\mathbf{H})}{\lambda_{\min}(\mathbf{H})}}$$

, where $\mathbf{H} = \frac{\partial^2 L}{\partial \boldsymbol{\epsilon}_\theta^2} \cdot \frac{\partial \boldsymbol{\epsilon}_\theta}{\partial \theta} \cdot \left(\frac{\partial \boldsymbol{\epsilon}_\theta}{\partial \theta}\right)^T + \frac{\partial L}{\partial \boldsymbol{\epsilon}_\theta} \cdot \frac{\partial^2 \boldsymbol{\epsilon}_\theta}{\partial \theta^2}$

Details in the paper!

Open-source: **microsoft/NeuralSpeech** @ GitHub

Samples: https://speechresearch.github.io/priorgrad/

DSAIL
Data Science & AI Laboratory

Microsoft