NAS-X: Neural Adaptive Smoothing via Twisting

Dieterich Lawson*, Michael Y. Li*, Scott W. Linderman

dieterichl@google.com,michaelyli@stanford.edu,scott.linderman@stanford.edu

Summary

Fitting **sequential latent variable models** is hard!

Why? High-variance gradients and discrete latents!

NAS-X = smoothing + reweighted wake sleep

- ❏ Low variance and low bias gradients.
- ❏ Versatile - handles discrete latents.
- ❏ Easy to train, minimal computational overhead.
- ❏ Significantly outperforms prior methods.

Reweighted Wake Sleep

Fisher's identity: Gradient of log marginal likelihood is a posterior expectation.

$$\nabla_\theta \log p_\theta(\mathbf{y}_{1:T}) = \mathbb{E}_{p_\theta(\mathbf{x}_{1:T} \mid \mathbf{y}_{1:T})}\left[\nabla_\theta \log p_\theta(\mathbf{x}_{1:T}, \mathbf{y}_{1:T})\right]$$

RWS estimates expectation via **importance sampling**.

$$\sum_{i=1}^{N} \overline{w}^{(i)} \nabla_\theta \log p_\theta(\mathbf{x}^{(i)}_{1:T}, \mathbf{y}_{1:T}), \quad \mathbf{x}^{(i)}_{1:T} \sim q_\phi(\mathbf{x}_{1:T} \mid \mathbf{y}_{1:T}).$$

$$\overline{w}^{(i)} \propto \frac{p_\theta(\mathbf{x}^{(i)}_{1:T}, \mathbf{y}_{1:T})}{q_\phi(\mathbf{x}^{(i)}_{1:T} \mid \mathbf{y}_{1:T})}$$

Estimating Posterior Expectations via Smoothing Sequential Monte Carlo

NAS-X estimates with **smoothing SMC with twists** r. Twists learned via density ratio estimation.

$$\sum_{t=1}^{T} \sum_{i=1}^{N} \overline{w}^{(i)}_t \nabla_\theta \log p_\theta(\mathbf{x}^{(i)}_t, \mathbf{y}_t \mid \mathbf{x}^{(i)}_{t-1})$$

$$\mathbf{x}^{1:N}_{1:T}, \overline{w}^{1:N}_{1:T} \leftarrow \mathrm{SMC}(\{p_\theta(\mathbf{x}_{1:t}, \mathbf{y}_{1:t}), q_\phi(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{y}_{t:T}), r_\psi(\mathbf{x}_t, \mathbf{y}_{t+1:T})\}_{t=1}^{T})$$

Under certain conditions, NAS-X has **unbiased** and **consistent** gradient estimates.

Algorithm

Algorithm 1: NAS-X

Procedure `NAS-X`$(\theta_0, \phi_0, \psi_0, \mathbf{y}_{1:T})$

$\theta \leftarrow \theta_0, \quad \phi \leftarrow \phi_0, \quad \psi \leftarrow \psi_0$

while *not converged* **do**

$\mathbf{x}^{1:N}_{1:T}, \overline{w}^{1:N}_{1:T} \leftarrow \mathrm{SMC}(\{p_\theta(\mathbf{x}_{1:t}, \mathbf{y}_{1:t}), q_\phi(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{y}_{t:T}), r_\psi(\mathbf{x}_t, \mathbf{y}_{t+1:T})\}_{t=1}^{T})$

$\Delta\theta = \sum_{t=1}^{T} \sum_{i=1}^{N} \overline{w}^{(i)}_t \nabla_\theta \log p_\theta(\mathbf{x}^{(i)}_t, \mathbf{y}_t \mid \mathbf{x}^{(i)}_{t-1})$

$\Delta\phi = -\sum_{t=1}^{T} \sum_{i=1}^{N} \overline{w}^{(i)}_t \nabla_\phi \log q_\phi(\mathbf{x}^{(i)}_t \mid \mathbf{x}^{(i)}_{t-1}, \mathbf{y}_{t:T})$

$\theta \leftarrow$ `grad-step`$(\theta, \Delta\theta)$

$\phi \leftarrow$ `grad-step`$(\phi, \Delta\phi)$

$\psi \leftarrow$ `twist-training`(θ, ψ)

end

return θ, ϕ, ψ

Procedure `twist-training`(θ, ψ_0)

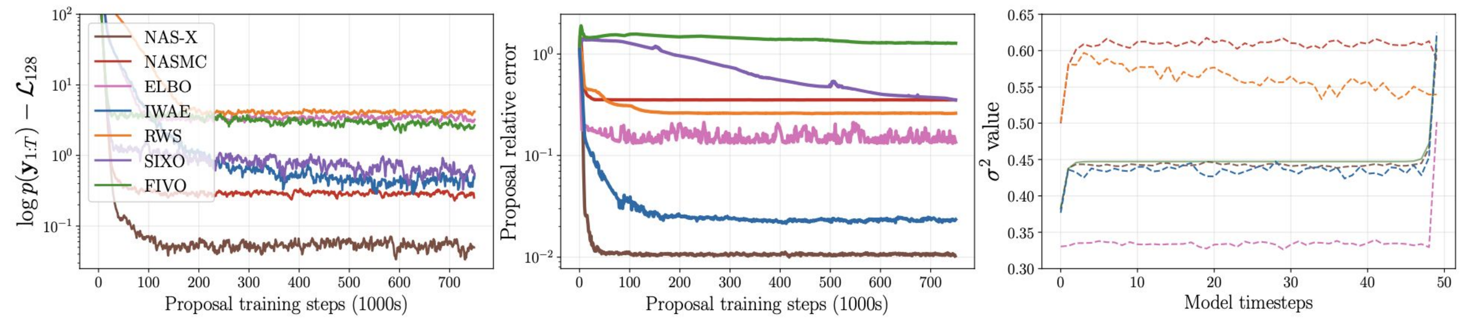
See Algorithm 2 in Appendix 8.3.

Paper

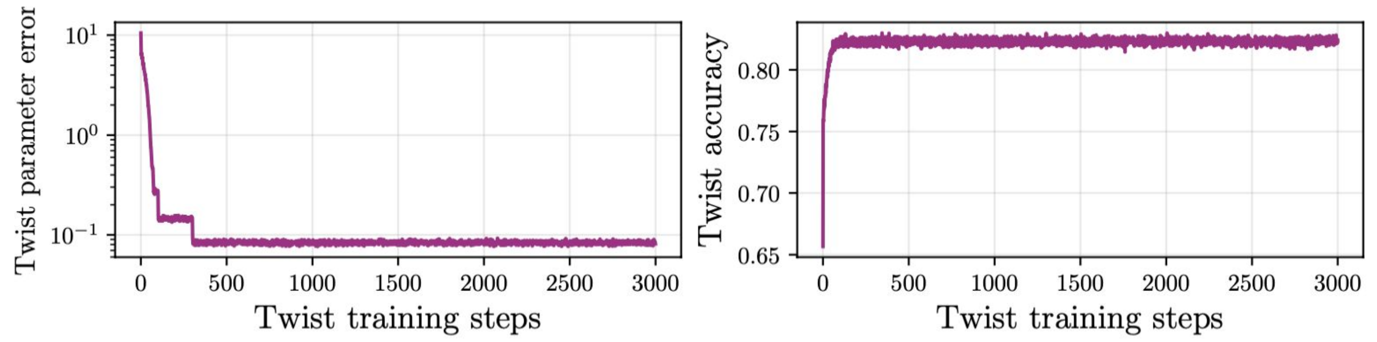
References

1. Dieterich Lawson, Allan Raventos, Andrew Warrington, and Scott Linderman. SIXO: Smoothing inference with twisted objectives. *Advances in Neural Information Processing Systems*, 2022.
2. Shixiang Gu, Zoubin Ghahramani, Richard E. Turner. Neural Adaptive Sequential Monte Carlo. *Advances in Neural Information Processing Systems*, 2015.
3. Alan L. Hodgkin and Andrew F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4):500, 1952.
4. Jörg Bornschein and Yoshua Bengio. Reweighted wake-sleep. arXiv preprint arXiv:1406.2751, 2014.

Linear Gaussian SSM

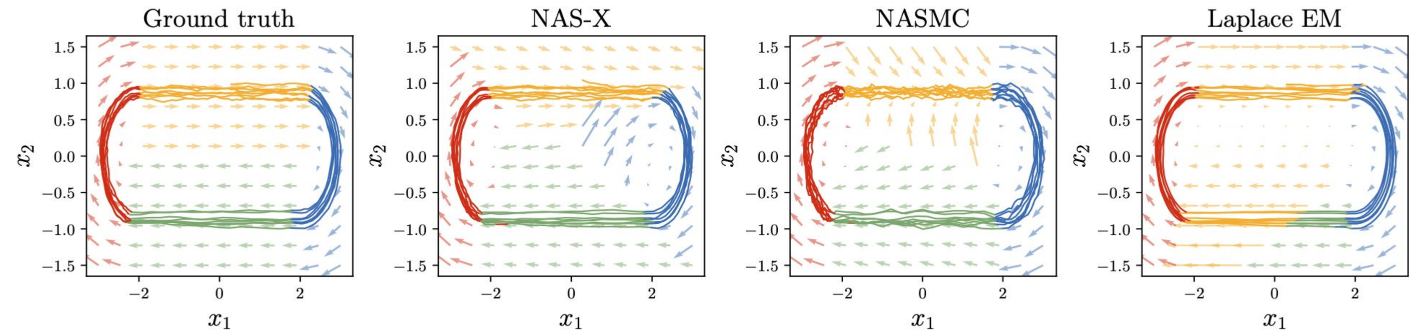


NAS-X has tighter variational bound, and recovers true posterior.



NAS-X recovers the true twist parameters.

Discrete latent variables



NAS-X can handle discrete latents.

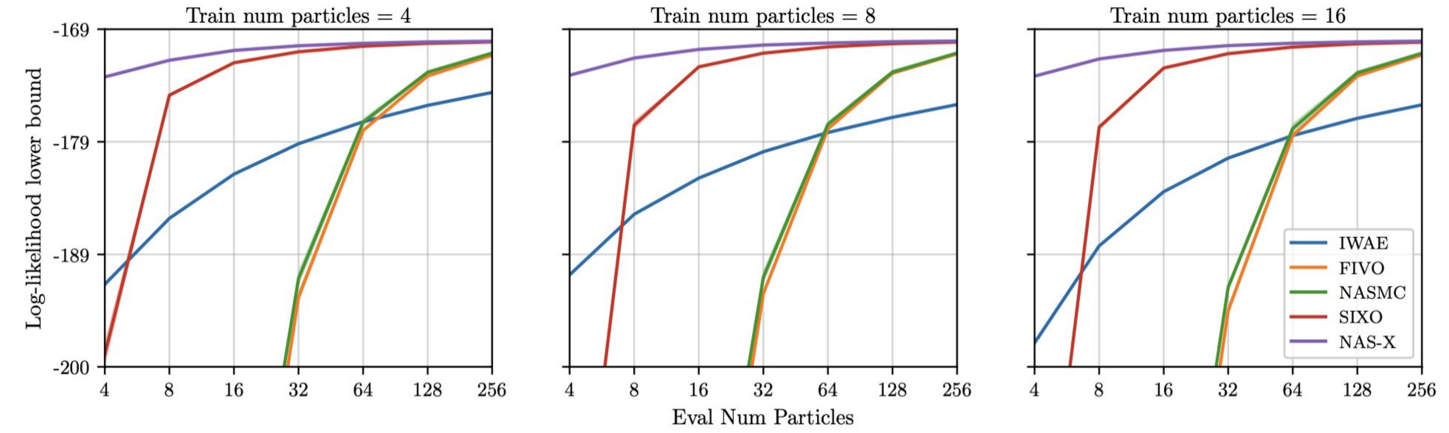
(b) Train $\mathcal{L}^{1024}_{\mathrm{BPF}}$ for rSLDS.

Method	$\sigma^2_O = 0.001$	$\sigma^2_O = 0.01$	$\sigma^2_O = 0.1$
NAS-X	$\mathbf{19.837 \pm 0.0234}$	$\mathbf{8.63 \pm 0.0015}$	-2.79 ± 0.0009
NASMC	19.834 ± 0.0018	8.53 ± 0.001	-2.874 ± 0.0007
Laplace EM	19.154 ± 0.057	8.54 ± 0.0039	$\mathbf{-2.765 \pm 0.0012}$
RWS	17.148 ± 0.087	6.314 ± 0.023	-5.78 ± 0.0026

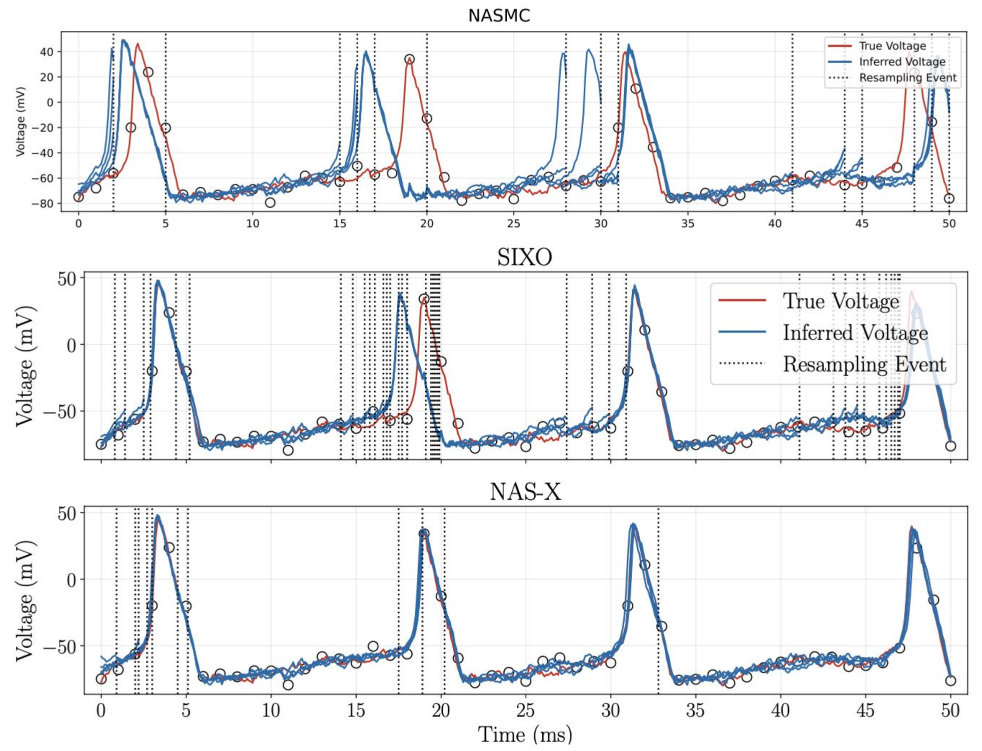
By smoothing, NAS-X learns better models than RWS methods.

Hodgkin Huxley model

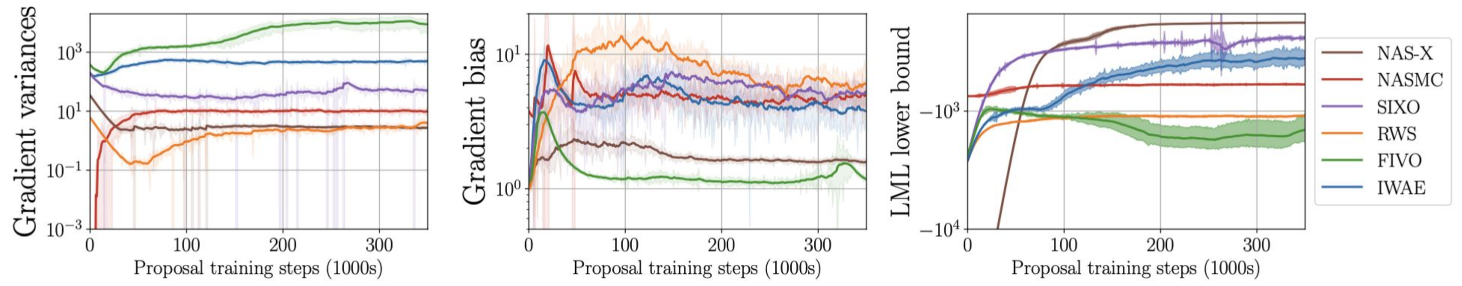
Hodgkin-Huxley, a mechanistic model of neural dynamics.



NAS-X is 4-64x more particle-efficient than prior methods.



NAS-X perfectly infers the latent voltage.



NAS-X has lower variance and lower bias gradients.