Variational Inference for Sequential Data with Future Likelihood Estimates

Geon-Hyeong Kim, Youngsoo Jang, Hongseok Yang, Kee-Eung Kim

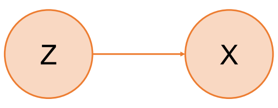
ICML 2020.

Presented by Shengyang Sun and Denny Wu.

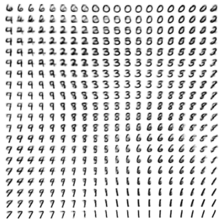
April 8, 2021

Deep Probabilistic Models

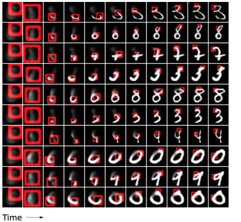
- Probabilistic latent variable models (LVMs) $p_\theta(x, z)$ describe high-dimensional structured data x using unobserved **latent variables** z.



- LVMs achieved remarkable successes when combined with deep learning[1].



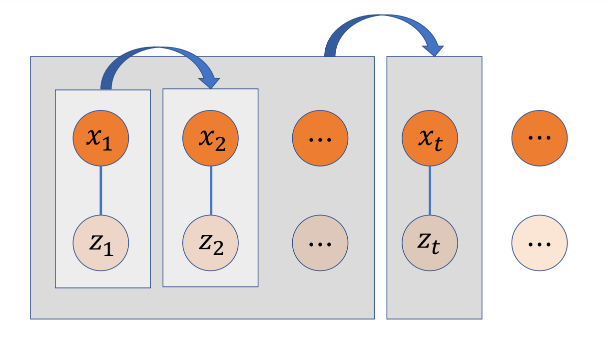
Variational auto-encoder.



Deep Recurrent Attentive Writer.

[1] *Kingma and Welling (2013)*; Gregor et al. (2015).

State-Space Models



When both x and z exhibit sequential structure, the joint density can be represented by a *state-space model*:

$$p_\theta(x, z) = p_\theta(x_1, z_1) \prod_{i=2}^{T} p_\theta(x_t, z_t | x_{1:t-1}, z_{1:t-1}),$$

where T is the length of the sequence, and $x_{i:j}$ denotes $(x_i, x_{i+1}, ..., x_j)$.

Variational Inference in State-Space Models

- A variational posterior is adopted for approximate inference,

$$q_\phi(z|x) = q_\phi(z_1|x)\prod_{t=2}^{T} q_\phi(z_t|z_{1:t-1}, x).$$

- q_ϕ can be optimized by maximizing the **Evidence Lower Bound (ELBO)**,

$$\mathcal{L}_{ELBO}(\theta, \phi; x) = \mathbb{E}_{q_\phi}\left[\log \frac{p_\theta(x, z)}{q_\phi(z|x)}\right] \leq \log p_\theta(x).$$

- **Importance-Weighted Auto-encoder (IWAE)** provides a tighter bound,

$$\mathcal{L}_{ELBO}(\theta, \phi; x) \leq \mathcal{L}_{IWAE}(\theta, \phi; x) \triangleq \mathbb{E}_{q_\phi}\left[\log\left(\frac{1}{N}\sum_{i=1}^{N} \underbrace{\frac{p_\theta(x, z^{(i)})}{q_\phi(z^{(i)}|x)}}_{w^{(i)}}\right)\right] \leq \log p_\theta(x),$$

by using *multiple particles* $N > 1$.

Gradient Estimators and the Baseline Method

Reparameterization Estimator.

- Typically lower variance.
- Only works for *differentiable* models.

Score Function Estimator.

- Applicable to discrete models.
- High gradient variance...

This work. Lower variance *score function estimator* for state-space models.

- **IWAE Gradient:** low-variance *path derivative* + high-variance *log derivative*.

$$\nabla_\phi \mathcal{L}_{IWAE}(\theta, \phi; x) = \nabla_\phi \mathbb{E}_{q_\phi}\left[\log\left(\frac{1}{N}\sum_{i=1}^{N} w_\phi^{(i)}\right)\right],$$

- Variance reduction is needed for the *log derivative* term.

The Log Derivative has High Variances

- **Intuition of High Variance.** *path derivative* has a *bounded* coefficient; the *path derivative* has an *unbounded* coefficient.

$$\text{path derivative:} \quad \sum_{i=1}^{N} \left[\frac{w_\phi^{(i)}}{\sum_{j=1}^{N} w_\phi^{(j)}} \right] \nabla_\phi \log q_\phi(z^{(i)}|x)$$

$$\text{log derivative:} \quad \sum_{i=1}^{N} \left[\log \left(\frac{1}{N} \sum_{j=1}^{N} w_\phi^{(j)} \right) \right] \nabla_\phi \log q_\phi(z^{(i)}|x)$$

- A **baseline** lowers the variance of the *log derivative* term.

$$\hat{g}_{high}(z^{(1:N)}; x) = \sum_{i=1}^{N} \left[\log \left(\frac{1}{N} \sum_{j=1}^{N} w_\phi^{(j)} \right) - B_i \right] \nabla_\phi \log q_\phi(z^{(i)}|x).$$

Constructing the Baseline

Question: How do we choose the baseline for variance reduction?

For a state-space model, the *log derivative* admits further decompositions,

$$\hat{g}_{high}(z^{(1:N)};x) = \sum_{i=1}^{N}\sum_{t=1}^{T}\left[\log\left(\frac{1}{N}\sum_{j=1}^{N} w^{(j)}\right) - B_{it}\right]\nabla_{\phi}\log q_{\phi}(z_t^{(i)}|z_{1:t-1}^{(i)},x),$$

$$w^{(j)} := \prod_{t=1}^{T}\frac{p_{\theta}(x_t, z_t^{(j)}|x_{1:t-1}, z_{1:t-1}^{(j)})}{q_{\phi}(z_t^{(j)}|z_{1:t-1}^{(j)},x)}.$$

Desiderata for the Baseline. The baseline B_{it} should

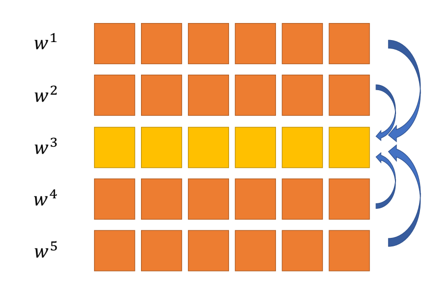
- Correlate with $\log\left(\frac{1}{N}\sum_{j=1}^{N} w^{(j)}\right)$.
- Be independent of $z_t^{(i)}$, so that $\mathbb{E}_{q(z)}[B\nabla\log q(z)] = 0$.

Prior Work: the VIMCO Estimator

Intuition: for each particle i and time t, we wish to construct B_{it} close to

$$\log\left(\frac{1}{N}\sum_{j=1}^{N} w^{(j)}\right) = \log\left(\underbrace{\frac{1}{N}\sum_{j\neq i}^{N} w^{(j)}}_{\text{Independent of } z_t} + \underbrace{w^{(i)}}_{\text{Need to replace}}\right).$$

Idea. Use *other particles* to "approximate" $w^{(i)}$.



Prior Work: the VIMCO Estimator

Intuition: for each particle i and time t, we wish to construct B_{it} close to

$$\log\left(\frac{1}{N}\sum_{j=1}^{N} w^{(j)}\right) = \log\left(\underbrace{\frac{1}{N}\sum_{j\neq i}^{N} w^{(j)}}_{\text{Independent of } z_t} + \underbrace{w^{(i)}}_{\text{Need to replace}} \right).$$

VIMCO: $w^{(i)}$ might be close to the *geometric mean* of other particles $w^{(\neg i)}$.

$$B_{it} = \log\left(\frac{1}{N}\sum_{j\neq i}^{N} w^{(j)} + \frac{1}{N}\prod_{j\neq i}^{N}\left(w^{(j)}\right)^{\frac{1}{N-1}}\right).$$

$$\text{Particle 1:} \quad w_1^{(1)} \times w_2^{(1)} \times \ldots w_{t-1}^{(1)} \times w_t^{(1)} \times w_{t+1}^{(1)} \times \ldots w_T^{(1)}$$

$$\ldots$$

$$\text{Particle i:} \quad \underbrace{w_1^{(i)} \times w_2^{(i)} \times \ldots w_{t-1}^{(i)} \times w_t^{(i)} \times w_{t+1}^{(i)} \times \ldots w_T^{(i)}}_{\text{Replaced by mean}\left(w^{(\neg i)}\right) \text{ in } B_{it}}$$

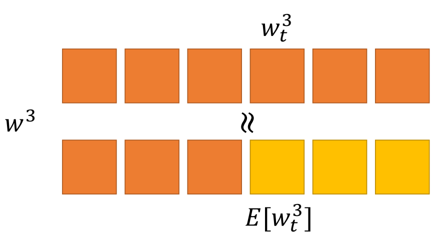
$$\ldots$$

The Future Likelihood Baseline

What is another way to replace the term $w^{(i)}$?

Idea: For each particle i and time t, $w^{(i)}$ should be close to $w_{1:t-1}^{(i)}\Gamma_{t-1}^{(i)}$, where

Γ is the **future likelihood function** defined as: $\Gamma_t^{(i)} \triangleq \mathbb{E}_{q(z_{t+\mathbf{1}:T}^{(i)}|z_{\mathbf{1}:t}^{(i)})}\left[w_{t+1:T}^{(i)}\right]$.



How do we estimate the future likelihood Γ?

- **Proposal:** parameterize Γ_t with a neural network.

The Future Likelihood Baseline

What is another way to replace the term $w^{(i)}$?

Idea: For each particle i and time t, $w^{(i)}$ should be close to $w^{(i)}_{1:t-1}\Gamma^{(i)}_{t-1}$, where

Γ is the **future likelihood function** defined as: $\Gamma^{(i)}_t \triangleq \mathbb{E}_{q(z^{(i)}_{t+\mathbf{1}:T}|z^{(i)}_{\mathbf{1}:t})}\left[w^{(j)}_{t+1:T}\right]$.

Proposed Method.

$$B_{it} := \log\left(\frac{1}{N}\sum_{j\neq i}^{N} w^{(j)} + \frac{1}{N} w^{(i)}_{1:t-1} \underbrace{\mathbb{E}_{q\left(z^{(i)}_{t:T}|z^{(i)}_{\mathbf{1}:t-\mathbf{1}}\right)}\left[w^{(i)}_{t:T}\right]}_{\text{Future Likelihood: } \Gamma^{(i)}_{t-\mathbf{1}}}\right),$$

$$\text{Particle 1:} \quad w^{(1)}_1 \times w^{(1)}_2 \times \dots w^{(1)}_{t-1} \times w^{(1)}_t \times w^{(1)}_{t+1} \times \dots w^{(1)}_T$$

$$\dots$$

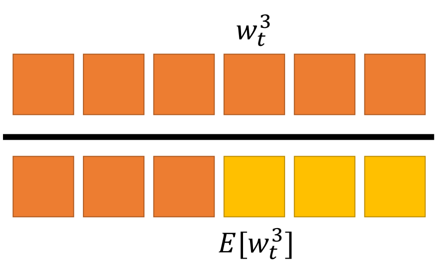
$$\text{Particle i:} \quad w^{(i)}_1 \times w^{(i)}_2 \times \dots w^{(i)}_{t-1} \times \underbrace{w^{(i)}_t \times w^{(i)}_{t+1} \times \dots w^{(i)}_T}_{\text{Replaced by } \Gamma^{(i)}_{t-\mathbf{1}} \text{ in } B_{it}}$$

$$\dots$$

Variational Inference with Future Likelihood Estimates (VIFLE)

- To sum up, the (unbiased) VIFLE gradient estimator is,

$$g^u_{VIFLE} = \sum_{i=1}^{N} \sum_{t=1}^{T} \left[\log \frac{\sum_{j \neq i}^{N} w^{(j)} + w^{(i)}_{1:t-1} w^{(i)}_{t:T}}{\sum_{j \neq i}^{N} w^{(j)} + w^{(i)}_{1:t-1} \Gamma^{(i)}_{t-1}} \right] \nabla_\phi \log q_\phi(z^{(i)}_t | z^{(i)}_{1:t-1}, x).$$

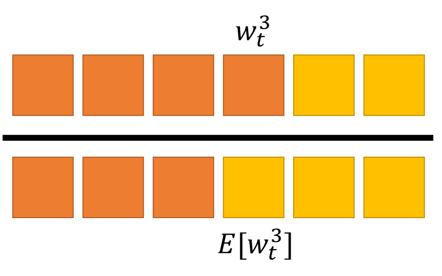


- High variances still persist due to the random variables $w^{(i)}_{t:T}$.

Variational Inference with Future Likelihood Estimates (VIFLE)

Proposal: introduce a surrogate objective:

$$g_{VIFLE} = \sum_{i=1}^{N} \sum_{t=1}^{T} \left[\log \frac{\sum_{j \neq i}^{N} w^{(j)} + w_{1:t}^{(i)} \Gamma_t^{(i)}}{\sum_{j \neq i}^{N} w^{(j)} + w_{1:t-1}^{(i)} \Gamma_{t-1}^{(i)}} \right] \nabla_\phi \log q_\phi(z_t^{(i)} | z_{1:t-1}^{(i)}, x).$$



- Now the denominator and enumerator differ by only one random variable $z_t^{(i)}$.
- g_{VIFLE} has lower variances but is **no longer unbiased**.

New Variational Lower Bound?

g_{VIFLE} is **biased**. What objective is it optimizing?

- Is it still a valid variational lower bound on $\log p_\theta(x)$?

Theorem (?). Let $\mathcal{L}_{VIFLE}(\theta, \phi; x)$ be the objective that g_{VIFLE} is optimizing, then

$$\log p_\theta(x) \geq \mathcal{L}_{IWAE}(\theta, \phi; x) \geq \mathcal{L}_{VIFLE}(\theta, \phi; x).$$

Is the theorem correct?

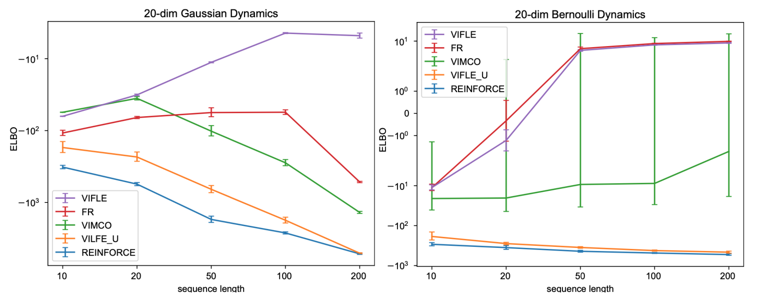
- **Sketch.** The proof of this theorem compares the IWAE bound to the VIFLE objective with *stop gradient*.
- It is meaningless to compare objectives that involve stop gradients. E.g.,

$$x * x \leq x * \mathrm{stopgrad}(x) + \mathrm{stopgrad}(x) * x,$$

- Yet LHS and RHS have the exact same gradient w.r.t. x.
- In reality $\mathcal{L}_{IWAE}(\theta, \phi; x) \leq \mathcal{L}_{VIFLE}(\theta, \phi; x)$...

Learning Simple Dynamical Systems.

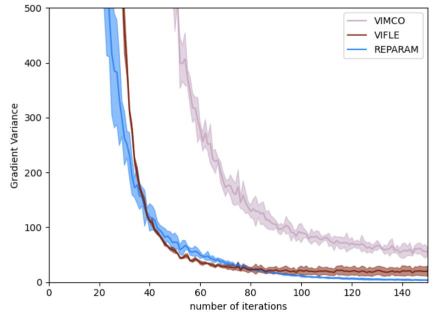
- Continuous Model: $z_t = Az_{t-1} + v_t, x_t = Bz_t + w_t$. $z, w \sim \mathcal{N}(0, \sigma^2 I_d)$.
- Discrete Model: $z_t = F(z_{t-1}), x_t = Az_t + \sin(10z_t) + w_t$. $z_0 \sim \text{Bern}(0.5)$.



- The (biased) VIFLE estimator consistently achieves good performance.

Learning Simple Dynamical Systems.

- Continuous Model: $z_t = Az_{t-1} + v_t, x_t = Bz_t + w_t$. $z, w \sim \mathcal{N}(0, \sigma^2 I_d)$.



- VIFLE outperforms VIMCO in terms of gradient variance, and is comparable to the reparameterization estimator.

Conclusion

Summary.

- Introduced a novel variance reduction method, the **future likelihood baseline**, for the IWAE objective in learning state-space models.
- Proposed a **biased** gradient estimator, g_{VIFLE}, to further reduce the gradient variance, and achieves strong empirical performance.

Limitation and Future Directions.

- What exactly is g_{VIFLE} optimizing?
  Is it still a valid lower bound on the log likelihood?
- Why does the *unbiased* VIFLE estimator perform very poorly?
  Is there a better way to parameterize the future likelihood?
- How does VIFLE compare to other *biased* estimators, e.g., Gumbel-Softmax?
- Can we apply similar baseline to objectives beyond IWAE?

Additional References

- Kingma and Welling 2014. *Auto-encoding variational bayes*.
- Gregor et al. 2015. *Draw: A recurrent neural network for image generation*.
- Burda et al. 2016. *Importance weighted autoencoders*.
- Mnih and Rezende 2016. *Variational inference for monte carlo objectives*.
- Maddison et al. 2017. *Filtering variational objectives*.
- Jang et al. 2017. *Categorical reparameterization with gumbel-softmax*.
- Maddison et al. 2017. *The concrete distribution: A continuous relaxation of discrete random variables*.

Appendix: VIFLE maximizes an Upper Bound of IWAE

Let $\mathcal{L}_{VIFLE}$ and $\mathcal{L}^u_{VILFE}$ be the objectives that g_{VIFLE} and g^u_{VIFLE} is optimizing, respectively. We have

$$
\begin{aligned}
&\mathcal{L}_{VIFLE}(z^{(1:N)}; x) - \mathcal{L}^u_{VIFLE}(z^{(1:N)}; x) \\
&= \mathbb{E}_{q_\phi}\left[\log \frac{\sum_{j\neq i}^N w^{(j)} + w^{(i)}_{1:t}\Gamma^{(i)}_t}{\sum_{j\neq i}^N w^{(j)} + w^{(i)}_{1:t-1}\Gamma^{(i)}_{t-1}}\right] - \mathbb{E}_{q_\phi}\left[\log \frac{\sum_{j\neq i}^N w^{(j)} + w^{(i)}_{1:t} w^{(i)}_{t+1:T}}{\sum_{j\neq i}^N w^{(j)} + w^{(i)}_{1:t-1}\Gamma^{(i)}_{t-1}}\right] \\
&= \mathbb{E}_{q_\phi}\left[\log\left(\sum_{j\neq i}^N w^{(j)} + w^{(i)}_{1:t}\Gamma^{(i)}_t\right)\right] - \mathbb{E}_{q_\phi}\left[\log\left(\sum_{j\neq i}^N w^{(j)} + w^{(i)}_{1:t} w^{(i)}_{t+1:T}\right)\right] \\
&\geq 0,
\end{aligned}
$$

The last inequality is due to Jensen's Inequality on $\Gamma^{(i)}_t = \mathbb{E}[w^{(i)}_{t+1:T}]$.

Appendix: Estimating the Future Likelihood Baseline

- Using random sampling to estimate $\Gamma_t^{(i)}$ is computationally expensive:

$$\Gamma_t = \mathbb{E}_{q(z_{t:T}|z_{\mathbf{1}:t-\mathbf{1}};x)}[w_{t:T}]$$

- Recall the recursive definition of the future likelihood function,

$$\Gamma_t(z_{1:t-1};x) = \mathbb{E}_{q(z_T|z_{\mathbf{1}:t-\mathbf{1}};x)}[w_t \Gamma_t(z_{1:t};x)],$$

$$w_{\mathbf{1}}^{(i)} \times w_{\mathbf{2}}^{(i)} \ldots \times \underbrace{w_t^{(i)} \times \overbrace{w_{t+\mathbf{1}}^{(i)} \ldots \times w_T^{(i)}}^{\Gamma_t^{(i)}}}_{\Gamma_{t-\mathbf{1}}^{(i)}}.$$

- **Proposed Method.** Learn a neural network to approximate the future likelihood. Parameterize $\hat{\Gamma}(z_{1:t-1};x)$ with a recurrent neural network. The following objective is minimized via *stochastic gradient descent*,

$$\min \sum_t \left(\hat{\Gamma}(z_{1:t-1};x) - \mathbb{E}_{q(z_T|z_{\mathbf{1}:t-\mathbf{1}};x)}[w_t \hat{\Gamma}(z_{1:t};x)]\right)^2.$$