Sample Average Approximation for Black-Box VI

Javier Burroni, Justin Domke, Daniel Sheldon

1. Setup

Focus: Variational inference for statistical models:

- hundreds of variables
- without data-subsampling

ELBO maximization:

- with reparameterizable distribution q_θ via $z_\theta(\cdot)$ and q_{base}

$$\max_{\theta\in\Theta} \mathscr{L}(\theta) = \max_{\theta\in\Theta} \mathbb{E}_{\epsilon\sim q_\theta}\left[\ln \frac{p(z_\theta(\epsilon), x)}{q_\theta(\epsilon)}\right]$$

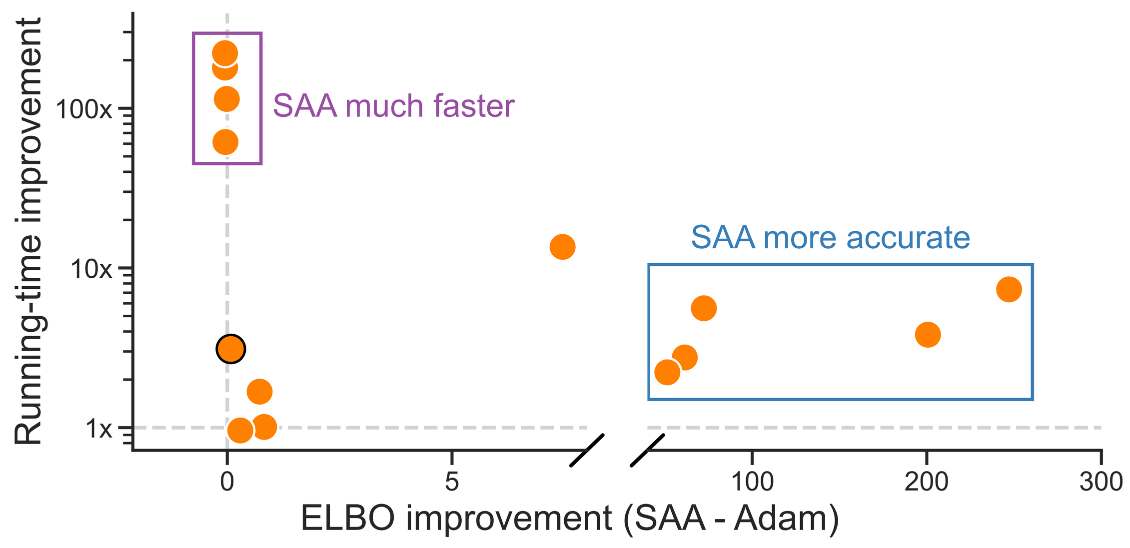
Usually solved with SGD:

- Hard to tune hyper-parameters
- Results highly dependent on choices

2. Contribution

We introduced **SAA for VI**:

- An alternative stochastic-optimization for BBVI
- Enhances both speed and quality of approximation



3. SAA

Take $\epsilon_1, \ldots, \epsilon_n \sim q_{\text{base}}$

Create deterministic optimization problem:

Solve $\max_{\theta\in\Theta} \widehat{\mathscr{L}}_\epsilon(\theta) = \max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n}\left[\ln \frac{p(z_\theta(\epsilon_i), x)}{q_\theta(z_\theta(\epsilon_i))}\right]$

Optimize with:

L-BFGS for search direction and **line search** for step-size

4. Sequence of SAA

Use sequence of sizes $n_1 < n_2 < \ldots$ to reduce Monte Carlo error

Algorithm 1 SAA for VI

1: **while not** converged$(\theta, \boldsymbol{\epsilon}_n)$ **do**
2: $\quad n \leftarrow 2n$
3: $\quad \boldsymbol{\epsilon}_n \leftarrow \epsilon_1, \ldots, \epsilon_n$, where $\epsilon_i \sim q_{\text{base}}$
4: $\quad \theta \leftarrow \text{Optimize}(\theta, \boldsymbol{\epsilon}_n)$
5: **return** $\theta^* \leftarrow \theta$

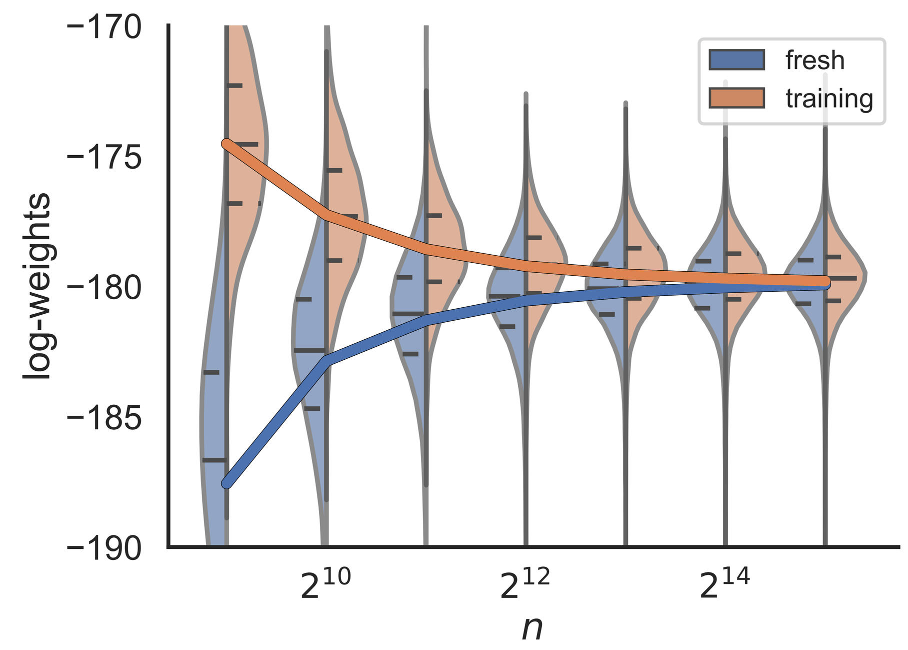
5. New convergence criterion

Compare distributions of log-weights.
Stop when training and testing cannot be distinguished

Algorithm 2 converged?

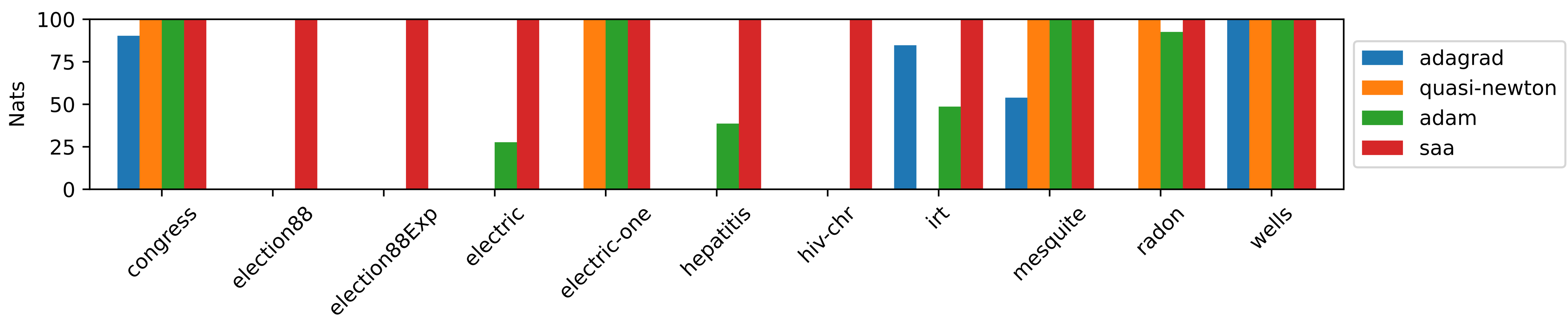
1: **Input:** $\theta, \boldsymbol{\epsilon}_n, t$ **Output:** True if converged
2: $\hat{\boldsymbol{\epsilon}}_{10k} \leftarrow \hat{\epsilon}_1, \ldots, \hat{\epsilon}_{10k}$, $\quad \hat{\epsilon}_i \sim q_{\text{base}}$
3: obj $\leftarrow$ mean$(v_\theta(\boldsymbol{\epsilon}_n))$
4: elbo $\leftarrow$ mean$(v_\theta(\hat{\boldsymbol{\epsilon}}_{10k}))$
$\quad \triangleright$ Statistically compare means:
5: $\text{p}_{\text{value}} \leftarrow \text{t\_test}(v_\theta(\boldsymbol{\epsilon}_n), v_\theta(\hat{\boldsymbol{\epsilon}}_{10k}))$
6: **return** $\text{p}_{\text{value}} > 0.01$

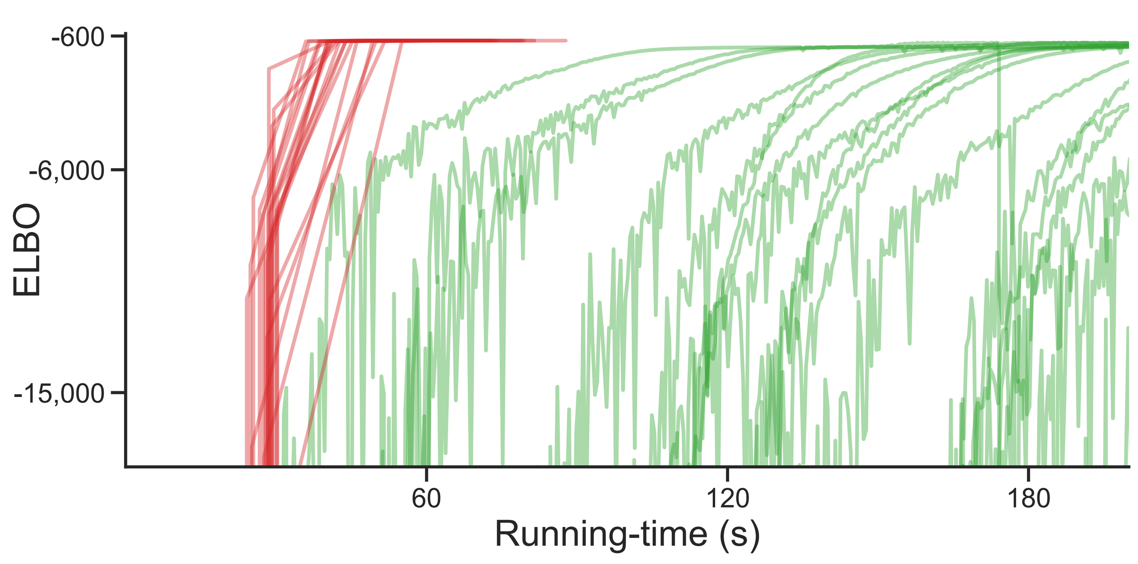
Training and **testing** log-weights distribution.
By increasing the sample size used for training we achieve a better approximation.



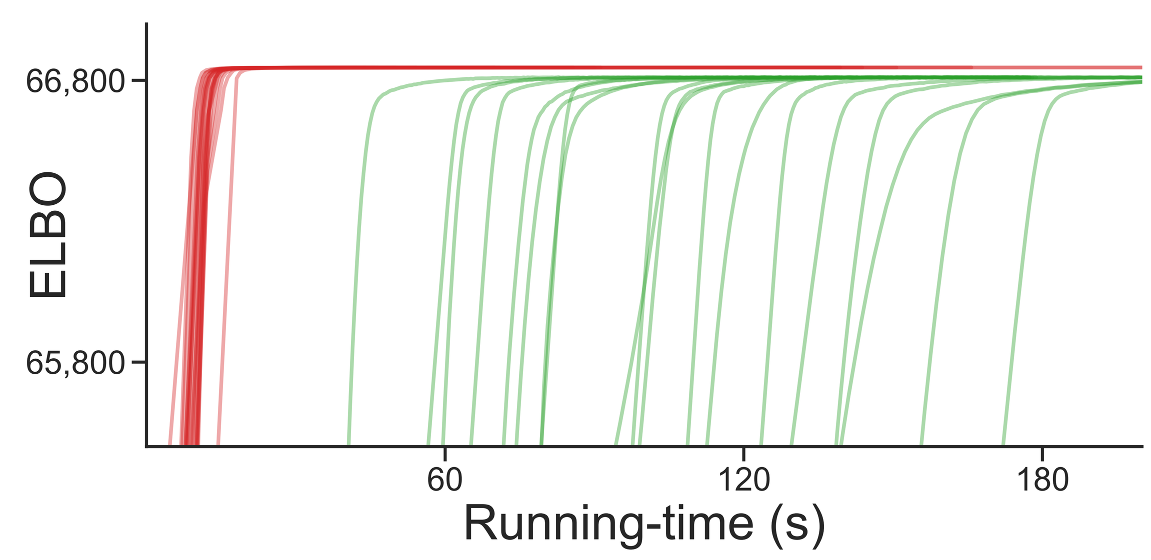
6. Experimental results

Stan models: ELBO comparison after training with dense Gaussian approximation.
For each model, ELBOs are shifted so the best method has value 100.





SAA for VI vs **Adam**:
Left: electric model ($D = 100$)
Right: Stochastic volatility ($D \approx 17k$)



UMass Amherst

contact: jburroni@cs.umass.edu