Structured Stochastic Gradient MCMC

Antonios Alexos[1]* Alex Boyd[2]* Stephan Mandt[1,2]

[1]Department of Computer Science, [2] Department of Statistics, University of California, Irvine

* Denotes equal contribution {aalexos,alexjb}@uci.edu

TL;DR

We propose a new non-parametric variational Langevin-type approximation that makes no parametric assumptions on the posterior distribution. It allows practitioners to impose arbitrary independence structures between parameters, resulting in faster training.

Preliminaries

- Stochastic gradient Markov chain Monte Carlo (SGMCMC) methods iteratively sample from the posterior distribution:

$$p(\theta|\mathcal{D}) \propto \exp\{-U(\theta)\} \text{ where } U(\theta) = -\log p(\theta, \mathcal{D}).$$

- A classic example is Stochastic Gradient Langevin Dynamics:

$$\theta^{(t+1)} = \theta^{(t)} - \frac{\epsilon_t}{2}\nabla_\theta \widehat{U}\big(\theta^{(t)}; \widetilde{\mathcal{D}}^{(t)}\big) + \xi_t, \qquad \xi_t \sim \mathcal{N}(0, \epsilon_t I),$$

$$\widehat{U}\big(\theta; \widetilde{\mathcal{D}}\big) = -\frac{|\mathcal{D}|}{|\widetilde{\mathcal{D}}|}\log p\big(\widetilde{\mathcal{D}}|\theta\big) - \log p(\theta), \qquad \mathbb{E}_{\widetilde{\mathcal{D}}}\big[\widehat{U}\big(\theta; \widetilde{\mathcal{D}}\big)\big] = U(\theta).$$

- Structured VI best approximates $p(\theta|\mathcal{D})$ with a restricted distribution $q(\theta) = \prod_{i=1}^{M} q_i(\theta_i)$ that assumes $\theta_i \perp \theta_j$ for $i \neq j$.
- This is casted as an optimization problem, minimizing the KL-divergence between them, yielding the following optimal results:

$$q^*(\theta) = \min_q D_{KL}(q(\theta)||p(\theta|\mathcal{D})) = \min_q \mathbb{E}_q\left[\log\frac{q(\theta)}{p(\theta|\mathcal{D})}\right]$$

$$= \exp\left\{\sum_{i=1}^{M} \mathbb{E}_{\tilde{\theta}_{\neg i}\sim q_{\neg i}}\big[\log\big(\theta_i, \tilde{\theta}_{\neg i}, \mathcal{D}\big)\big]\right\}$$

- Typically, parametric assumptions are placed on each q_i and Coordinate Ascent is performed on the summands of $q^*(\theta)$.

Structured SGMCMC

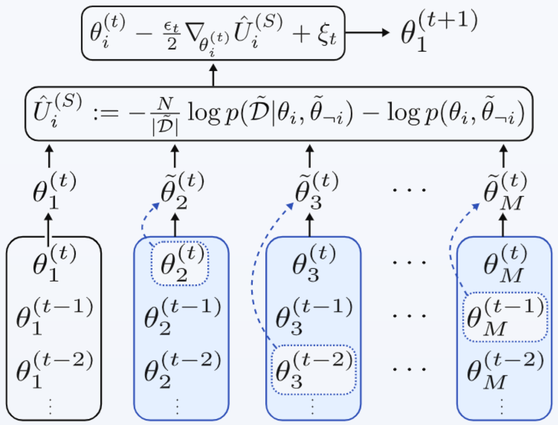
- Instead, we want to sample directly from $q^*(\theta)$ via SGMCMC and avoid any parametric assumptions on $q(\theta)$.
- To do so, we perform SGMCMC with $\widehat{U}\big(\theta; \widetilde{\mathcal{D}}\big)$ replaced by:

$$\widehat{U}^{(S)}\big(\theta; \widetilde{\mathcal{D}}\big) = \sum_{i=1}^{M} \mathbb{E}_{\tilde{\theta}_{\neg i}\sim q_{\neg i}} \widehat{U}\big(\{\theta_i, \tilde{\theta}_{\neg i}\}; \widetilde{\mathcal{D}}\big),$$

$$= \sum_{i=1}^{M} \mathbb{E}_{\tilde{\theta}_{\neg i}\sim q_{\neg i}}\left[-\frac{|\mathcal{D}|}{|\widetilde{\mathcal{D}}|}\log p\big(\widetilde{\mathcal{D}}|\theta_i, \tilde{\theta}_{\neg i}\big) - \log p\big(\theta_i, \tilde{\theta}_{\neg i}\big)\right].$$

- When generating $\theta^{(t+1)}$, we approximate $\mathbb{E}_{\tilde{\theta}_{\neg i}\sim q_{\neg i}}$ with MC samples: $\tilde{\theta}_{\neg i} \sim \prod_{j\neq i}\{\theta_j^{(1)}, \theta_j^{(2)}, \ldots, \theta_j^{(t)}\}$

Structured SGMCMC Visualized



Structured Dropout SGMCMC

- Generating a sample for θ with Structured SGMCMC requires evaluating $\widehat{U}^{(S)}(\theta; \widetilde{\mathcal{D}})$ which requires $\mathcal{O}(M)$ model forward passes.
- To avoid having computation scaled by the number of parameter groups, we develop a further approximation to sampling from $q^*(\theta)$.
- First, we recognize that:

$$\widehat{U}^{(S)}\big(\theta; \widetilde{\mathcal{D}}\big) \equiv M\mathbb{E}_{r\sim \text{Cat}(M^{-1},\ldots,M^{-1})}\mathbb{E}_{\tilde{\theta}\sim q}\widehat{U}\big(r\theta + (1-r)\tilde{\theta}; \widetilde{\mathcal{D}}\big)$$

- Using MC samples for the outer expectation breaks scaling issues but leads to sparse gradients (⇒ not every parameter group being sampled).
- Proposed method replaces $r \sim \text{Cat}(M^{-1}, \ldots, M^{-1})$ with $r \sim p_{\text{mask}}$ where $r \in [0,1]^M$ and $\sum_i r_i > 0$, which yields a new approximate energy function:

$$\widehat{U}^{(S_d)}\big(\theta; \widetilde{\mathcal{D}}\big) \equiv \frac{M}{\mathbb{E}_{r\sim p_{\text{mask}}}\sum_i r_i}\mathbb{E}_{r\sim p_{\text{mask}}}\mathbb{E}_{\tilde{\theta}\sim q}\widehat{U}\big(r\theta + (1-r)\tilde{\theta}; \widetilde{\mathcal{D}}\big)$$

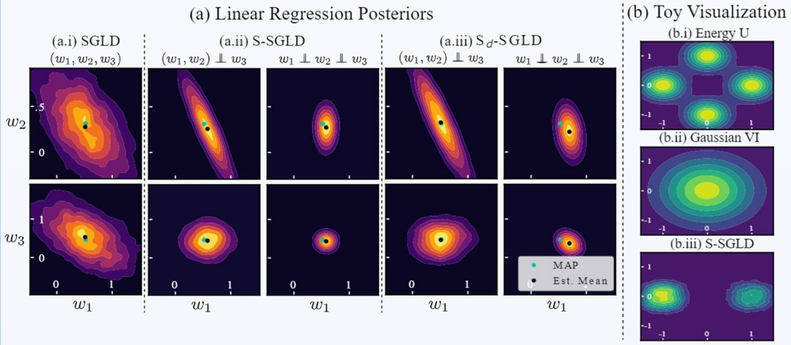
- Example masking distributions:

$$p_{\text{mask}}(r) = \prod_{i=1}^{M}\text{Unif}\,(r_i; (0,1)) \qquad p_{\text{mask}}(r) = \prod_{i=1}^{M}\text{Bern}\,(r_i; \pi)\mathbf{1}(\exists_i r_i = 1)$$

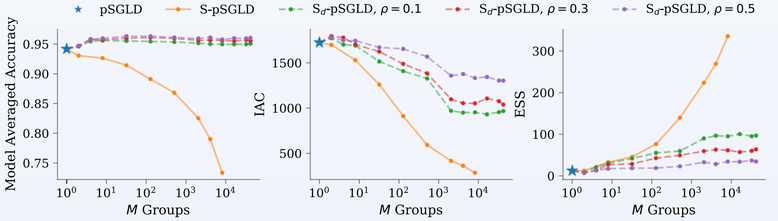
- Structured Dropout SGMCMC can be seen as an interpolation between Structured and Unstructured SGMCMC:

$$r \xrightarrow{d} \text{Cat}(M^{-1}, \ldots, M^{-1}) \Rightarrow U^{(S_d)} \rightarrow U^{(S)} \qquad r \xrightarrow{d} \{1\}^M \Rightarrow U^{(S_d)} \rightarrow U$$

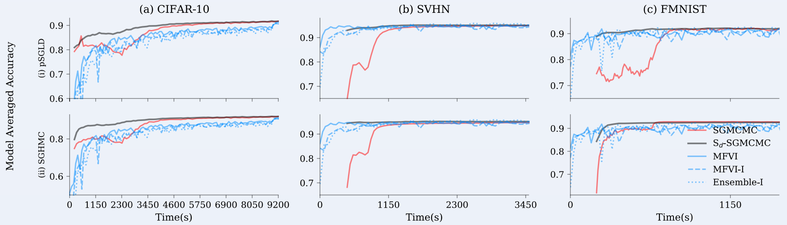
Experimental Visualizations



Independence Amount Investigation



- Shown above are the accuracy, integrated autocorrelation time (IAC), and effective sample size (ESS) as a function of the number of independent parameter groups for a NN trained on MNIST.
- More broken correlations equals improvement on IAC, ESS.
- Performance of S_d-SGMCMC improves compared to S-SGMCMC.



- Accuracy vs. wall-clock time for two base SGMCMC methods, S_d-SGLD and some VI baselines. S_d-SGMCMC methods converge much faster and sometimes they outperform the VI baselines.