Approx. Inference via Sampling (Contd): MCMC with Gradients, Recent Advances

CS772A: Probabilistic Machine Learning

Piyush Rai

Plan for today

- Some other aspects of MCMC
- MCMC with gradient
- Some other recent advances

Sampling Methods: Label Switching Issue

- Suppose we are given samples $\boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)}, \ldots, \boldsymbol{Z}^{(S)}$ from the posterior $p(\boldsymbol{Z}|\boldsymbol{X})$
- We can't always simply "average" them to get the "posterior mean" $\bar{\boldsymbol{Z}}$
- Why: Non-identifiability of latent vars in models with multiple equival. posterior modes
- Example: In clustering via GMM, the likelihood is invariant to how we label clusters
  - What we call cluster 1 in one sample may be cluster 2 in the next sample
  - Say, in GMM, $\boldsymbol{z}_n^{(1)} = [1,0]$ and $\boldsymbol{z}_n^{(2)} = [0,1]$, both may imply the same
  - Averaging will give $\bar{\boldsymbol{z}}_n = [0.5,0.5]$, which is incorrect

One sample may be from near one of the modes and the other may be from near the other mode

- Quantities not affected by permutations of dims of $\boldsymbol{Z}$ can be safely averaged
  - E.g., probability that two points belong to the same cluster (e.g., in GMM)
  - Predicting the mean of an entry r_{ij} in matrix factorization $\frac{1}{S}\sum_{s=1}^{S} \boldsymbol{u}_i^{(s)\top} \boldsymbol{v}_j^{(s)}$

Changes in order of entries in these $K \times 1$ vectors across different samples doesn't affect the inner product

MCMC: Some Practical Aspects

- Choice of proposal distribution is important
  - For MH sampling, Gaussian proposal is popular when $\boldsymbol{z}$ is continuous, e.g.,

$$q\left(\boldsymbol{z}\middle|\boldsymbol{z}^{(\ell-1)}\right) = \mathcal{N}\left(\boldsymbol{z}\middle|\boldsymbol{z}^{(\ell-1)}, \mathbf{H}\right)$$

Hessian at the MAP of the target distribution

Change at each iter

  - Other options: Mixture of proposal distributions, data-driven or adaptive proposals

- Autocorrelation. Can show that when approximating $f^* = \mathbb{E}[f]$ using $\{\boldsymbol{Z}^{(s)}\}_{s=1}^{S}$

$\bar{f} = \frac{1}{S}\sum_{s=1}^{S} f_s$

Monte Carlo assumes uncorrelated samples

Value of f using s^{th} MCMC sample

$$\text{var}_{MCMC}[\bar{f}] = \text{var}_{MC}[\bar{f}] + \frac{1}{S^2}\sum_{s\neq t}\mathbb{E}[(f_s - f^*)(f_t - f^*)]$$

Basically measures what fractions of the total samples are uncorrelated. Want it to be close to 1

$$\text{Effective Sample Size (ESS)} = \frac{\text{var}_{MC}[f]}{\text{var}_{MCMC}[f]}$$

- Autocorrelation function (ACF) at lag $\boldsymbol{t}$: $\rho_t = \frac{\frac{1}{S-t}\sum_{s=1}^{S-t}(f_s-\bar{f})(f_{s+t}-\bar{f})}{\frac{1}{S-1}\sum_{s=1}^{S}(f_s-\bar{f})^2}$

Lower is better

- Multiple Chains: Run multiple chains, take union of generated samples

Coming Up Next

- Avoiding the random-walk behavior of MCMC
  - Using gradient information of the posterior
- Scalable MCMC methods

Using Gradients in MCMC: Langevin Dynamics

- MCMC uses a random-walk based proposal to generate the next sample, e.g.,

$$\theta^{(t)} \sim \mathcal{N}(\theta^{(t-1)}, \eta_t)$$

And then accept/reject (MH)

Will use θ to denote all the unknowns

Can use automatic differentiation methods for this

- Langevin dynamics: Use (unnormalized) posterior's gradient info in the proposal as

$$\theta^* = \theta^{(t-1)} + \frac{\eta_t}{2} \nabla_\theta [\log p(\mathcal{D}|\theta) + \log p(\theta)]\Big|_{\theta^{(t-1)}}$$

Likelihood

Prior

Move towards the mode of the posterior (like finding MAP est)

$$\theta^{(t)} \sim \mathcal{N}(\theta^*, \eta_t)$$

And then accept/reject (MH)

η set s.t. acceptance rate is around 0.6

Same as doing a gradient ascent step towards the posterior and injecting noise $\epsilon_t \sim \mathcal{N}(0, \eta_t)$. Noise ensures we aren't stuck at the MAP solution

Using gradient info in the proposal helps us move faster towards high-prob regions

- Note that the above is equivalent to

$$\theta^{(t)} = \theta^{(t-1)} + \frac{\eta_t}{2} \nabla_\theta [\log p(\mathcal{D}|\theta) + \log p(\theta)]\Big|_{\theta^{(t-1)}} + \epsilon_t$$

Helps also incorporate the curvature info of the posterior

If gradient is pre-multiplied by a preconditioner matrix $M(\theta_t)$: Simplified Manifold MALA

And then accept/reject (MH)

Known as Metropolis-Adjusted Langevin Algorithm (MALA)

One option to use for $M(\theta_t)$ is the second derivative of the unnorm. posterior

- After some waiting period T_0, iterates $\{\theta^{(t)}\}_{t=T_0+1}^{T_0+S}$ are MCMC samples from $p(\theta|\mathcal{D})$

"Bayesian Learning via Stochastic Gradient Langevin Dynamics" by Welling and Teh (2011)

Langevin Dynamics: A Closer Look

- Is generating MCMC samples really as easy as computing MAP?
- Recall the form of Langevin Dynamics updates

And then accept/reject (MH)

$$\theta^{(t)} = \theta^{(t-1)} + \frac{\eta_t}{2}\nabla_\theta[\log p(\mathcal{D}|\theta) + \log p(\theta)]\big|_{\theta^{(t-1)}} + \epsilon_t$$

Same as our target posterior

- Equivalent to discretization of an SDE with equilibrium distribution $\propto \exp(\log p(\mathcal{D}, \theta))$

Above update is its discretiization

Note that this is continuous time

$$d\theta_t = -\nabla L(\theta_t)dt + \sqrt{2}dB_t$$

where $L(\theta_t) = -\log p(\mathcal{D}, \theta_t)$ and $(B_t)_{t\geq 0}$ is Brownian motion s.t. ΔB_t are i.i.d. Gaussian r.v.s

- Discretization introduces some error which is corrected by MH accept/reject step
- Note: As learning rate η_t decreases, discretization error also decreases and rejection rate tends to zero
- Note: Gradient computations require all the data (thus slow)
  - Solution: Use stochastic gradients - Stochastic Gradient Langevin Dynamics (SGLD)

Stochastic Gradient Langevin Dynamics (SGLD)

- An "online" MCMC method: Langevin Dynamics with minibatches to compute gradients
- Given minibatch $\mathcal{D}_t = \{x_{t1}, x_{t2}, \ldots, x_{tN_t}\}$, the (stochastic) Langevin dynamics update:

$$\theta^* = \theta^{(t-1)} + \frac{\eta_t}{2}\nabla_\theta \left[\frac{N}{|\mathcal{D}_t|}\sum_{n=1}^{N_t} \log p(x_{tn}|\theta) + \log p(\theta)\right]$$

Almost as fast as doing SGD updates ☺

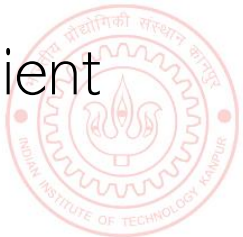
And then accept/reject (MH)

$$\theta^{(t)} \sim \mathcal{N}(\theta^*, \eta_t)$$

- Choice of the learning rate is important. For convergence, $\eta_t = a(b+t)^{-\kappa}$
  - Switching to constant learning rates (after a few iterations) often helps convergence
- As η_t becomes very very small, acceptance prob. becomes close to 1

No need for accept/reject (MH)

- Recent flurry of work on this topic (see "Bayesian Learning via Stochastic Gradient Langevin Dynamics" by Welling and Teh (2011) and follow-up works)



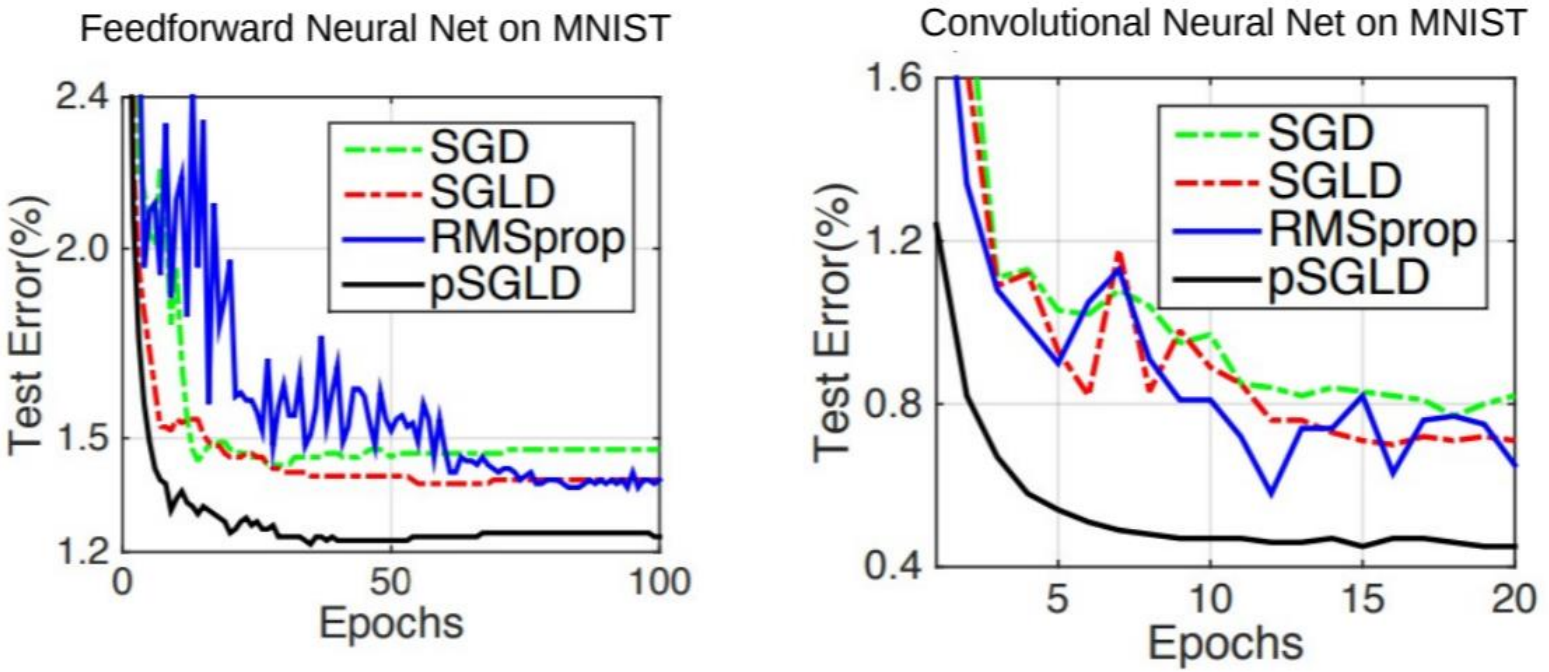
Improvements to SGLD

- The basic SGLD, although fairly simple, has many limitations, e.g.
  - Exhibits slow convergence and mixing. Uses same learning rate η_t in all dimensions of θ
  - Doesn't apply to models where θ is constrained (e.g., non-neg or prob. vector)
  - Needs to the model to be differentiable (since it needs $\nabla_\theta \log p(\mathcal{D}, \theta)$)
- A lot of recent work on improving the basic SGLD to handle such limitations
- Introducing the curvature information in the gradients, e.g.,
  - Bayesian Posterior Sampling via Stochastic Gradient Fisher Scoring (Ahn et al, 2012), and Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks (Li et al, 2016)
  - These methods use a preconditioner matrix in the learning rate to improve convergence
  - This also allows different amounts of updates in different dimensions
- SLGD in Riemannian space to handle constrained variables
  - Stoch. Grad. Riemannian Langevin Dynamics on the Probability Simplex (Patterson and Teh, 2013)

Based on reparametrizing the constrained variables to make them unconstrainted

Applications of SGLD

- Popular for Bayesian neural networks and other complex Bayesian models
- Reason: SGLD = backprop based updates + Gaussian noise



(Figure: Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks (Li et al, 2016))

Other Recent "SGD-inspired" Sampling Algorithms

- Run SGD and use SGD iterates $\theta_1, \theta_2, \ldots, \theta_T$ to construct a Gaussian approximation
- Recently Maddox et al (2019) proposed an idea using stochastic weight avging (SWA)

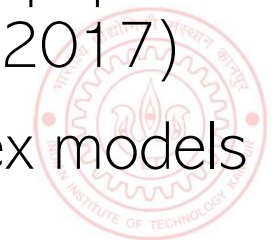
$$\theta_{SWA} = \frac{1}{T}\sum_{t=1}^{T}\theta_t$$

$$\bar{\theta^2} = \frac{1}{T}\sum_{t=1}^{T}\theta_t^2, \quad \Sigma_{\text{diag}} = \text{diag}(\bar{\theta^2} - \theta_{SWA}^2)$$

$$p(\theta|\mathcal{D}) \approx \mathcal{N}(\theta_{SWA}, \Sigma_{\text{diag}})$$

Approach known as SWA-Gaussian (SWAG)

- If we want full cov., we can use a low-rank approx. of $\boldsymbol{\Sigma}$ (see Maddox et al for details)
- Reason it works: SGD is asymptotically Normal under certain conditions
- For a more detailed theory of SGD and MCMC, may also refer to this very nice paper: Stochastic Gradient Descent as Approximate Bayesian Inference (Mandt et al, 2017)
- Such algos can give not too accurate but very fast posterior approx for complex models



A Simple Baseline for Bayesian Uncertainty in Deep Learning, Maddox et al (2019)

Hamiltonian/Hybrid Monte Carlo (HMC)

- HMC (Neal, 1996) is an "auxiliary variable sampler" and incorporates gradient info
- Uses the idea of simulating a Hamiltonian Dynamics of a physical system
- Consider the target posterior $p(\theta|\mathcal{D}) \propto \exp(-U(\theta))$
- Think of θ as "position" then $U(\theta) = -\log p(\mathcal{D}|\theta)p(\theta)$ is like "potential energy"
- Let's introduce an auxiliary variable - the momentum $\boldsymbol{r}$ of the system
- Can now define a joint distribution over the position and momentum as

$$p(\theta, \boldsymbol{r}) \propto \exp\left(-U(\theta) - \frac{1}{2}\boldsymbol{r}^\top M^{-1}\boldsymbol{r}\right) = p(\theta|\mathcal{D})p(\boldsymbol{r})$$

- The total energy (potential + kinetic) or the Hamiltonian of the system

Constant w.r.t. time

$$H(\theta, \boldsymbol{r}) = U(\theta) + \frac{1}{2}\boldsymbol{r}^\top M^{-1}\boldsymbol{r} = U(\theta) + K(\boldsymbol{r})$$

- Given a sample $(\boldsymbol{\theta}, \boldsymbol{r})$ from $p(\boldsymbol{\theta}, \boldsymbol{r})$, ignoring $\boldsymbol{r}$, θ will be a sample from $p(\boldsymbol{\theta}|\mathcal{D})$

Generating Samples in HMC

- Given an initial $(\boldsymbol{\theta}, \boldsymbol{r})$, Hamiltonian Dynamics defines how $(\boldsymbol{\theta}, \boldsymbol{r})$ changes w.r.t. time t

$$\frac{\partial \theta}{\partial t} = \frac{\partial H}{\partial \boldsymbol{r}} = \frac{\partial K}{\partial \boldsymbol{r}}$$
$$\frac{\partial \boldsymbol{r}}{\partial t} = -\frac{\partial H}{\partial \theta} = -\frac{\partial U}{\partial \theta}$$

$$\left(H(\theta, \boldsymbol{r}) = U(\theta) + \tfrac{1}{2}\boldsymbol{r}^\top M^{-1}\boldsymbol{r} = U(\theta) + K(\boldsymbol{r}) \right)$$

- We can use these equations to update $(\boldsymbol{\theta}, \boldsymbol{r}) \rightarrow (\boldsymbol{\theta}^*, \boldsymbol{r}^*)$ by discretizing time

Reason: Getting analytical solutions for the above requires integrals which is in general intractable

- For $s = 1{:}S$, sample as follows
  - Initialize $\theta_0 = \theta^{(s-1)}$, $\boldsymbol{r}_* \sim \mathcal{N}(0, \mathbf{I})$ and $\boldsymbol{r}_0 = \boldsymbol{r}_* - \frac{\rho}{2}\frac{\partial U}{\partial \theta}|_{\theta_0}$
  - Do L "leapfrog" steps with learning rates $\rho_\ell = \rho$ for $\ell < L$ and $\rho_L = \rho/2$
    - For $\ell = 1{:}L$

$$\theta_\ell = \theta_{\ell-1} + \rho \frac{\partial K}{\partial \boldsymbol{r}}|_{\boldsymbol{r}_{\ell-1}}$$
$$\boldsymbol{r}_\ell = \boldsymbol{r}_{\ell-1} - \rho_\ell \frac{\partial U}{\partial \theta}|_{\theta_\ell}$$

L usually set to 5 and learning rate tuned to make acceptance rate around 90%

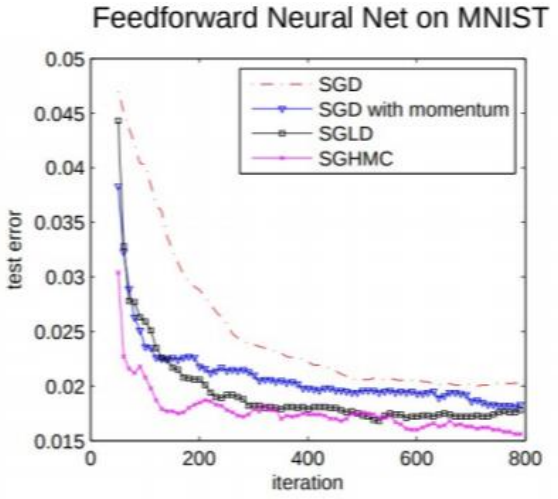
  - Perform MH accept/reject test on $(\boldsymbol{\theta}_L, \boldsymbol{r}_L)$. If accepted $\theta^{(s)} = \theta_L$

A single sample generated by taking L steps

- The momentum forces exploring different regions instead of getting driven to regions where the MAP solution is

HMC in Practice

- HMC typically has very low rejection rate (that too, primarily due to discretization error)
- Performance can be sensitive to L (no. of leapfrog steps) and step-sizes, tuning hard
- A lot of renewed interest in HMC (you may check out NUTS - No U-turn Sampler – doesn't require setting L)
  - Prob. Prog. packages e.g., Tensorflow Prob., Stan, etc, contain implementations of HMC
- Can also do HMC on minibatches (Stochastic Gradient HMC - Chen et al, 2014)
- An illustration: SGHMC vs other methods on MNIST classification (Bayesian neural net)



(Figure: Stochastic Gradient Hamiltonian Monte Carlo (Chen et al, 2014))

Parallel/Distributed MCMC

- Suppose our goal is to compute the posterior of $\theta \in \mathbb{R}^D$ (assuming N is very large)

$$p(\theta|\mathbf{X}) \propto p(\theta)p(\mathbf{X}|\theta) = p(\theta)\prod_{n=1}^{N} p(\boldsymbol{x}_n|\theta)$$

- Suppose we have J machines with data partitioned as $\mathbf{X} = \{\mathbf{X}^{(j)}\}_{j=1}^{J}$
- Let's assume that the posterior $p(\theta|\mathbf{X})$ factorizes as

$$p(\theta|\mathbf{X}) = \prod_{i=1}^{J} p^{(j)}(\theta|\mathbf{X}^{(j)})$$

- Here $p^{(j)}(\theta|\mathbf{X}^{(j)}) \propto p(\theta)^{1/J}\prod_{\boldsymbol{x}_n \in \mathbf{X}^{(j)}} p(\boldsymbol{x}_n|\theta)$ is known as the "subset posterior"
- Assume the j^{th} machine generates T MCMC samples $\{\theta_{j,t}\}_{t=1}^{T}$
- We need a way to combine these subset posteriors using a "consensus"

$$\hat{\theta}_1, \ldots, \hat{\theta}_T = \texttt{CONSENSUSSAMPLES}(\{\theta_{j,1}, \ldots, \theta_{j,T}\}_{j=1}^{J})$$

Parallel/Distributed MCMC

- Many ways to compute the consensus samples. Let's look at two of them
- Approach 1: Weighted Average: $\hat{\theta}_t = \sum_{j=1}^{J} W_j \theta_{j,t}$ where W_j can be learned as follows
  - Assuming Gaussian likelihood and Gaussian prior

$$\begin{aligned} \bar{\Sigma}_j &= \text{sample covariance of } \{\theta_{j,1}, \ldots, \theta_{j,T}\} \\ \Sigma &= (\Sigma_0^{-1} + \sum_{j=1}^{J} \bar{\Sigma}_j^{-1})^{-1} \quad (\Sigma_0 \text{ is the prior's covariance}) \\ W_j &= \Sigma(\Sigma_0^{-1}/J + \bar{\Sigma}_j^{-1}) \end{aligned}$$

These approaches can also be used to make VI parallel/distributed

- Approach 2: Fit J Gaussians, one for each $\{\theta_{j,t}\}_{t=1}^{T}$ and take their product

$$\begin{aligned} \bar{\mu}_j &= \text{sample mean of } \{\theta_{j,1}, \ldots, \theta_{j,T}\}, \quad \bar{\Sigma}_j = \text{sample covariance of } \{\theta_{j,1}, \ldots, \theta_{j,T}\} \\ \hat{\Sigma}_J &= (\sum_{j=1}^{J} \bar{\Sigma}_j^{-1})^{-1}, \quad \hat{\mu}_J = \hat{\Sigma}_J(\sum_{j=1}^{J} \bar{\Sigma}_j^{-1}\bar{\mu}_j) \quad \text{(cov and mean of prod. of Gaussians)} \\ \hat{\theta}_t &\sim \mathcal{N}(\hat{\mu}_J, \hat{\Sigma}_J), t = 1, \ldots, T \quad \text{(the final consensus samples)} \end{aligned}$$

- For detailed proofs and other approaches, may refer to the reference below

Patterns of Scalable Bayesian Inference (Angelino et al, 2016)

Approximate Inference: VI vs Sampling

- VI approximates a posterior distribution $p(\boldsymbol{Z}|\boldsymbol{X})$ by another distribution $q(\boldsymbol{Z}|\boldsymbol{\phi})$
- Sampling uses $\boldsymbol{S}$ samples $\boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)}, \ldots, \boldsymbol{Z}^{(S)}$ to approximate $p(\boldsymbol{Z}|\boldsymbol{X})$
- Sampling can be used within VI (ELBO approx using Monte-Carlo)
- In terms of "comparison" between VI and sampling, a few things to be noted
  - Convergence: VI only has local convergence, sampling (in theory) can give exact posterior
  - Storage: Sampling based approx needs to storage all samples, VI only needs var. params $\boldsymbol{\phi}$
  - Prediction Cost: Sampling always requires Monte-Carlo avging for posterior predictive; with VI, sometimes we can get closed form posterior predictive

PPD if using sampling: $$p(x_*|X) = \int p(x_*|Z)p(Z|X)dZ \approx \frac{1}{S}\sum_{s=1}^{S} p\left(x_*|Z^{(s)}\right)$$

PPD if using VI: $$p(x_*|X) = \int p(x_*|Z)p(Z|X)dZ \approx \int p(x_*|Z)q(Z|\phi)dZ$$

Compressing the S samples into something more compact

- There is some work on "compressing" sampling-based approximations*

*"Compact approximations to Bayesian predictive distributions" by Snelson and Ghaharamani, 2005; and "Bayesian Dark Knowledge" by Korattikara et al, 2015

Inference Methods: Summary

- MLE/MAP: Straightforward for differentiable models (can even use automatic diff.)
- Conjugate models with one "main" parameter: Straightforward posterior updates
- MLE-II/MAP-II: Often useful for estimating the hyperparameters
- EM: If we want to do MLE/MAP for models with latent variables
  - Very general algorithm, can also be made online
  - Used when we want point estimates for some unknowns and posterior over others
  - Can use it for hyperparameter estimation as well
  - Often better than using direct gradient methods
- VI and sampling methods can be used to get full posterior for complex models
  - Quite easy if we have local conjugacy (VI has closed form updates, Gibbs sampler is easy to derive)
  - In other cases, we have general VI with Monte-Carlo gradients, MH sampling
  - MCMC can also make use of gradient info (LD/SGLD)
- For large-scale problems, online/distributed VI/MCMC, or SGD based posterior approx