Doubly Stochastic Variational Bayes for non-Conjugate Inference (ICML 2014)

Michalis K. Titsias, Miguel Lázaro-Gredilla

Discussion led by Yan Kaganovsky
Duke University

Goal

In variational Bayesian inference we maximize the ELBO

$$\max_{\boldsymbol{\lambda}} \mathcal{F}(\boldsymbol{\lambda}) = \max_{\boldsymbol{\lambda}} \int q(\boldsymbol{\theta}|\boldsymbol{\lambda}) \log \frac{p(\mathbf{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta}|\boldsymbol{\lambda})} d\boldsymbol{\theta} =$$
$$= \max_{\boldsymbol{\lambda}} \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\lambda})} \log p(\mathbf{y}, \boldsymbol{\theta}) + \mathcal{H}_q(\boldsymbol{\lambda})$$

which is equivalent to minimizing $\mathrm{KL}[q(\boldsymbol{\theta}|\boldsymbol{\lambda})||p(\boldsymbol{\theta}|\mathbf{y})]$

- Often $\mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\lambda})} \log p(\mathbf{y}, \boldsymbol{\theta})$ and its gradient $\nabla_{\boldsymbol{\lambda}}$ do not have a closed-form expression.
- Paisley et al. 2012 suggested a stochastic search method to circumvent this difficulty.
- The current paper proposes another method which is more efficient and algorithmically simpler.

Theory

- Consider the random vector $\mathbf{z} \in \mathbb{R}^D$ with pdf $\phi(\mathbf{z})$.
- Assume $\phi(\mathbf{z})$ exists in standard form with zero mean and scale parameters set to 1. For example:
    - standard Normal distribution
    - standard t distribution
    - product of standard logistic distributions
- Assume $\phi(\mathbf{z})$ permits straightforward simulation of independent samples.
- We can change the mean and correlations by applying an invertible transformation

$$\boldsymbol{\theta} = \mathbf{C}\mathbf{z} + \boldsymbol{\mu}$$

where $\mathbf{C}$ is a lower triangular psd matrix.

Theory

The pdf for $\boldsymbol{\theta}$ takes the form

$$q(\boldsymbol{\theta}|\boldsymbol{\mu}, \mathbf{C}) = \frac{1}{|\mathbf{C}|}\phi(\mathbf{C}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}))$$

This will be the variational approximation to the posterior (generally correlated with free parameters $\boldsymbol{\mu}, \mathbf{C}$)

The authors focus on $\phi(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$ for which

$$q(\boldsymbol{\theta}|\boldsymbol{\mu}, \mathbf{C}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \mathbf{C}\mathbf{C}^T)$$

As in Challis & Barber (2011).

Theory

Define $g(\boldsymbol{\theta}) = p(\mathbf{y}, \boldsymbol{\theta})$. The ELBO is given by

$$\mathcal{F}(\boldsymbol{\mu}, \mathbf{C}) = \int q(\boldsymbol{\theta}|\boldsymbol{\mu}, \mathbf{C}) \log \frac{g(\boldsymbol{\theta})}{q(\boldsymbol{\theta}|\boldsymbol{\mu}, \mathbf{C})} d\boldsymbol{\theta}$$

By variable transformation back to $\mathbf{z} = \mathbf{C}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})$ we get

$$\begin{aligned}\mathcal{F}(\boldsymbol{\mu}, \mathbf{C}) &= \int \phi(\mathbf{z}) \log \frac{g(\mathbf{C}\mathbf{z} + \boldsymbol{\mu})|\mathbf{C}|}{\phi(\mathbf{z})} d\mathbf{z} \\ &= \mathbb{E}_{\phi(\mathbf{z})}[\log g(\mathbf{C}\mathbf{z} + \boldsymbol{\mu})] + \underbrace{\log|\mathbf{C}|}_{\sum_k \log(C_{kk})} + \underbrace{\mathcal{H}_\phi}_{Const(\boldsymbol{\mu},\mathbf{C})}\end{aligned}$$

Theory

To fit the variational distribution we maximize $\mathcal{F}$. We need to compute

$$\nabla_{\boldsymbol{\mu}} \mathcal{F}(\boldsymbol{\mu}, \mathbf{C}) = \mathbb{E}_{\phi(\mathbf{z})}[\nabla_{\boldsymbol{\mu}} \log g(\mathbf{C}\mathbf{z} + \boldsymbol{\mu})]$$
$$\nabla_{\mathbf{C}} \mathcal{F}(\boldsymbol{\mu}, \mathbf{C}) = \mathbb{E}_{\phi(\mathbf{z})}[\nabla_{\mathbf{C}} \log g(\mathbf{C}\mathbf{z} + \boldsymbol{\mu})] + \Delta_{\mathbf{C}}$$

where $\Delta_{\mathbf{C}} = \text{diag}(1/C_{11}, ..., 1/C_{DD})$. Alternatively, going back to $\boldsymbol{\theta} = \mathbf{C}\mathbf{z} + \boldsymbol{\mu}$ and using the chain rule we obtain

$$\nabla_{\boldsymbol{\mu}} \mathcal{F}(\boldsymbol{\mu}, \mathbf{C}) = \mathbb{E}_{q(\boldsymbol{\theta})}[\nabla_{\boldsymbol{\theta}} \log g(\boldsymbol{\theta})]$$
$$\nabla_{\mathbf{C}} \mathcal{F}(\boldsymbol{\mu}, \mathbf{C}) = \mathbb{E}_{q(\boldsymbol{\theta})}[\nabla_{\boldsymbol{\theta}} \log g(\boldsymbol{\theta}) \times \underbrace{(\boldsymbol{\theta} - \boldsymbol{\mu})^T \mathbf{C}^{-T}}_{\mathbf{z}^T}] + \Delta_{\mathbf{C}}$$

where in the latter we take only the lower triangular part.

Doubly Stochastic Gradient Ascent

Use an unbiased Monte Carlo estimator for the expectation

$$\nabla_{\boldsymbol{\mu}}\mathcal{F} = \mathbb{E}_{q(\boldsymbol{\theta})}[\underbrace{\nabla_{\boldsymbol{\theta}} \log g(\boldsymbol{\theta})}_{f(\boldsymbol{\theta})}] \approx \frac{1}{S}\sum_{s=1}^{S} f(\boldsymbol{\theta}^s), \qquad \boldsymbol{\theta}^s \sim^{i.i.d} q(\boldsymbol{\theta})$$

Based on the theory of stochastic approximations (Robbins & Monro, 1951) we use a sample instead of the full gradient

$$\nabla_{\boldsymbol{\mu}}\mathcal{F} \rightarrow \nabla_{\boldsymbol{\theta}} \log g(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^s}$$

Algorithm

Algorithm 1 Doubly stochastic variational inference

Input: ϕ, $\mathbf{y}$, $\boldsymbol{\theta}$, $\nabla \log g$.
Initialise $\boldsymbol{\mu}^{(0)}, C^{(0)}$, $t = 0$.
repeat
  $t = t + 1$;
  $\mathbf{z} \sim \phi(\mathbf{z})$;
  $\boldsymbol{\theta}^{(t-1)} = C^{(t-1)}\mathbf{z} + \boldsymbol{\mu}^{(t-1)}$;
  $\boldsymbol{\mu}^{(t)} = \boldsymbol{\mu}^{(t-1)} + \rho_t \nabla_{\boldsymbol{\theta}} \log g(\boldsymbol{\theta}^{(t-1)})$;
  $C^{(t)} = C^{(t-1)} + \rho_t \left(\nabla_{\boldsymbol{\theta}} \log g(\boldsymbol{\theta}^{(t-1)}) \times \mathbf{z}^T + \Delta_{C^{(t-1)}}\right)$;
until convergence criterion is met.

The learning rate (step sizes) $\{\rho_t\}$ must satisfy $\sum_t \rho_t = \infty$ and $\sum_t \rho_t^2 < \infty$ to guarantee convergence to a local maximum (or global when $\mathcal{F}$ is concave).

To appreciate the proposed Algorithm, let us review prior art:

- Straightforward integration (Opper & Archambeau 2009, Challis & Barber 2011,2013)
- Variational Bayesian Inference with Stochastic Search (Paisley, Blei, Jordan 2012)

Straightforward Integration

Gaussian Approximation $q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}^T)$ (Opper & Archambeau 2009, Challis & Barber 2011,2013)

$$\mathcal{F} = \int \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \log[g(\boldsymbol{\theta})] d\boldsymbol{\theta} + \frac{1}{2} \sum_j \log C_{jj}$$

Straightforward Integration

Gaussian Approximation $q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}^T)$ (Opper & Archambeau 2009, Challis & Barber 2011,2013)

$$\nabla_{\boldsymbol{\mu}} \mathcal{F} = \int \nabla_{\boldsymbol{\mu}} \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \log[g(\boldsymbol{\theta})] d\boldsymbol{\theta} =$$

$$= \int \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \log[g(\boldsymbol{\theta})] \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}) d\boldsymbol{\theta}$$

Straightforward Integration

Gaussian Approximation $q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}^T)$ (Opper & Archambeau 2009, Challis & Barber 2011,2013)

if $g(\boldsymbol{\theta}) = \prod_{n=1}^{N} g_n(\mathbf{h}_n^T \boldsymbol{\theta}) \prod_j p(\theta_j)$ then

$$\mathcal{F} = \sum_{n=1}^{N} \int \mathcal{N}(z; 0, 1) g_n(\mathbf{h}_n^T \boldsymbol{\mu} + z \mathbf{h}_n^T \boldsymbol{\Sigma} \mathbf{h}_n) dz + \\ + \sum_j \int \mathcal{N}(\theta_j; \mu_j, \Sigma_{jj}) \log p(\theta_j) d\theta_j + \frac{1}{2} \sum_j \log C_{jj}$$

Reduces to 1D integrals

Review of Paisley, Blei, Jordan 2012

$$\nabla_{\boldsymbol{\lambda}} \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\lambda})}[f(\boldsymbol{\theta})] = \nabla_{\boldsymbol{\lambda}} \int f(\boldsymbol{\theta}) q(\boldsymbol{\theta}|\boldsymbol{\lambda}) \boldsymbol{\theta}$$

$$= \int f(\boldsymbol{\theta}) \underbrace{\nabla_{\boldsymbol{\lambda}} q(\boldsymbol{\theta}|\boldsymbol{\lambda})}\, d\boldsymbol{\theta} = \int f(\boldsymbol{\theta}) \underbrace{{\color{red} q(\boldsymbol{\theta}|\boldsymbol{\lambda})} \nabla_{\boldsymbol{\lambda}} \log q(\boldsymbol{\theta}|\boldsymbol{\lambda})}\, d\boldsymbol{\theta}$$

$$= \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\lambda})}[f(\boldsymbol{\theta}) \nabla_{\boldsymbol{\lambda}} \log q(\boldsymbol{\theta}|\boldsymbol{\lambda})] \approx \frac{1}{S} \sum_{s=1}^{S} f(\boldsymbol{\theta}^s) \nabla_{\boldsymbol{\lambda}} \log q(\boldsymbol{\theta}^s|\boldsymbol{\lambda})$$

where $\boldsymbol{\theta}^s \sim^{i.i.d} q(\boldsymbol{\theta}|\boldsymbol{\lambda})$.

- f can be any distribution and $\log q$ must be smooth
- Paisley's method has high variance $\rightarrow$ requires control variates (complicates the algorithm)

Comparison

Method	Relevant Models	Approximation to the Posterior	Speed	Computation
Integration	Separable w.r.t. data pts (to reduce integrals to 1D)	Gaussian	Fastest	• numerical integration might be required for function and gradient evaluations • simple when integrals are analytic
Doubly Stochastic	smooth log priors and log likelihoods	Also non-Gaussian	Moderate	Only gradient of the joint pdf
Paisley	Any	• Smooth log posterior • easy to draw	Slowest (due to slow convergence)	• gradient of the approximate posterior • Compute sample variances and covariances (control variates)

Bayesian logistic regression on the Pima diabetes dataset.
Integration: 16 likelihood evaluations with L-BFGS.
DSVI: 500 evaluations ($\times 3$ more time)

Illustrative Convergence Analysis

Very simple example:
The model $f(\boldsymbol{\theta}) = \log p(\mathbf{y}, \boldsymbol{\theta}) = \log(\text{const} \times \mathcal{N}(\boldsymbol{\theta}; \mathbf{m}, \mathbf{I}))$
Approximate posterior $q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \mathbf{I})$ so $\boldsymbol{\mu}^* = \mathbf{m}$

Illustrative Convergence Analysis

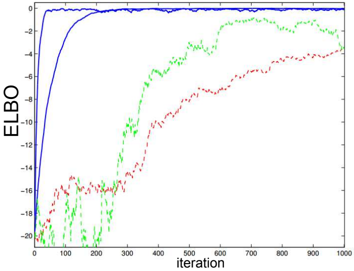
Very simple example:
The model $f(\boldsymbol{\theta}) = \log p(\mathbf{y}, \boldsymbol{\theta}) = \log(\text{const} \times \mathcal{N}(\boldsymbol{\theta}; \mathbf{m}, \mathbf{I}))$
Approximate posterior $q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \mathbf{I})$ so $\boldsymbol{\mu}^* = \mathbf{m}$

DSVI uses $\boldsymbol{\mu}^{(t)} = \boldsymbol{\mu}^{(t-1)} + \rho_t(\mathbf{m} - \boldsymbol{\theta}^s)$
Paisley/Direct uses $\boldsymbol{\mu}^{(t)} = \boldsymbol{\mu}^{(t-1)} + \rho_t[f(\boldsymbol{\theta}^s)(\boldsymbol{\theta}^s - \boldsymbol{\mu}^{(t-1)})]$
where $\boldsymbol{\theta}^s \sim \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}^{(t-1)}, \mathbf{I})$

Illustrative Convergence Analysis

Very simple example:
The model $f(\boldsymbol{\theta}) = \log p(\mathbf{y}, \boldsymbol{\theta}) = \log(\text{const} \times \mathcal{N}(\boldsymbol{\theta}; \mathbf{m}, \mathbf{I}))$
Approximate posterior $q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \mathbf{I})$ so $\boldsymbol{\mu}^* = \mathbf{m}$

DSVI uses $\boldsymbol{\mu}^{(t)} = \boldsymbol{\mu}^{(t-1)} + \rho_t(\mathbf{m} - \boldsymbol{\theta}^s)$
Paisley/Direct uses $\boldsymbol{\mu}^{(t)} = \boldsymbol{\mu}^{(t-1)} + \rho_t[f(\boldsymbol{\theta}^s)(\boldsymbol{\theta}^s - \boldsymbol{\mu}^{(t-1)})]$
where $\boldsymbol{\theta}^s \sim \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}^{(t-1)}, \mathbf{I})$



Variable Selection in Logistic Regression

The authors propose a DSVI-ARD algorithm.
$q(\boldsymbol{\theta}; \boldsymbol{\mu}, \mathbf{c}) = \prod_{d=1}^{D} q(\theta_d; \mu_d, c_d)$ with prior $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \mathbf{0}, \boldsymbol{\Lambda})$ with $\boldsymbol{\Lambda} = \text{diag}(\ell_1^2, ..., \ell_D^2)$

$$\mathcal{F}(\boldsymbol{\mu}, \mathbf{C}, \boldsymbol{\Lambda}) = \mathbb{E}_{\phi(\mathbf{z})}[\log \tilde{g}(\mathbf{c} \circ \mathbf{z} + \boldsymbol{\mu})] + \frac{1}{2}\sum_{d=1}^{D} \log c_d^2 + \\ - \frac{1}{2}\sum_{d=1}^{D} \log \ell_d^2 - \frac{1}{2}\sum_{d=1}^{D} \frac{c_d^2 + \mu_d^2}{\ell_d^2} + \frac{D}{2}$$

The point estimate for the hyperparameters is $(\ell_d^2)^* = c_d^2 + \mu_d^2$. Substituting this we obtain

$$\mathbb{E}_{\phi(\mathbf{z})}[\log \tilde{g}(\mathbf{c} \circ \mathbf{z} + \boldsymbol{\mu})] + \frac{1}{2}\sum_{d=1}^{D} \log(c_d^2) - \frac{1}{2}\sum_{d=1}^{D} \log(c_d^2 + \mu_d^2)$$

Variable Selection in Logistic Regression

Table 1. Size and number of features of each cancer data set.

Data set	#Train	#Test	D
Colon	42	20	2,000
Leukemia	38	34	7,129
Breast	38	4	7,129

Table 2. Train and test errors for the three cancer datasets and for each method: CONCAV is the original DSVI algorithm with a fixed prior, whereas ARD is the feature-selection version.

Problem	Train Error	Test Error
Colon (ARD)	0/42	1/20
Colon (CONCAV)	0/42	0/20
Leukemia (ARD)	0/38	3/34
Leukemia (CONCAV)	0/38	12/34
Breast (ARD)	0/38	2/4
Breast (CONCAV)	0/38	0/4

Table 3. Size and sparsity level of each large-scale data set.

Data set	#Train	#Test	D	#Nonzeros
a9a	32,561	16,281	123	451,592
rcv1	20,242	677,399	47,236	49,556,258
Epsilon	400,000	100,000	2,000	800,000,000

Table 4. Test error rates for DSVI-ARD and ℓ_1-logistic regression on three large-scale data sets.

Data set	DSVI ARD	Log. Reg.	λ
a9a	0.1507	0.1500	2
rcv1	0.0414	0.0420	4
Epsilon	0.1014	0.1011	0.5