# Conjugate-Computation Variational Inference (CVI)

Mohammad Emtiyaz Khan
RIKEN Center for Advanced Intelligence Project (AIP)

Joint work with W Lin (AI-Stats 2017)

# Uncertainty Estimation is Computationally Challenging

Bayes' rule

$$p(\mathbf{z}|\mathbf{y}) = \frac{p(\mathbf{y}, \mathbf{z})}{\int p(\mathbf{y}, \mathbf{z}) d\mathbf{z}}$$

Exact computation of the integral is difficult.

# Variational Inference

Integration to Optimization

$$\log \int p(\mathbf{y}, \mathbf{z}) d\mathbf{z}$$

$$\geq \max_{\lambda} \; \mathbb{E}_q \left[ \log \frac{p(\mathbf{y}, \mathbf{z})}{q(\mathbf{z}|\lambda)} \right]$$

High-dimensional "intractable" lower bound optimization

# Stochastic Gradient Descent

$$\lambda_{t+1} = \lambda_t + \beta_t \, \frac{\partial \mathcal{L}(\lambda_t)}{\partial \lambda}$$

| | SGD | CVI |
|---|---|---|
| General? | ✓ | ✓ |
| Scalable? | ✓ | ✓ |
| Computationally Efficient? | ✗ | ✓ |
| Modular? | ✗ | ✓ |
| Independent of parameterization? | ✗ | ✓ |

# Conjugate-Computation VI (CVI)

- Two modifications to SGD $\lambda_{t+1} = \lambda_t + \beta_t \dfrac{\partial \mathcal{L}(\lambda_t)}{\partial \lambda}$
  - Optimize in the mean-parameter space.
  - Change the geometry to KL divergence (natural gradients)
- Natural gradient step can be expressed as an "inference in a conjugate model".
  - Logistic Regression to Linear Regression
  - GP classification to GP Regression
  - Advanced Topic model to LDA
- In general, mean-field using message passing
- Structured inference on deep models.

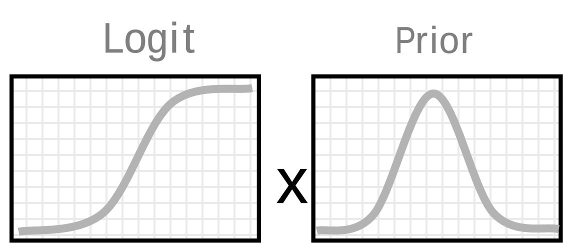
# Example of Non-conjugate models

Bayes' rule

$$p(\mathbf{z}|\mathbf{y}) = \frac{p(\mathbf{y}, \mathbf{z})}{\int p(\mathbf{y}, \mathbf{z}) d\mathbf{z}}$$

Gaussian Process classification (GPC)

$$\int \left[ \prod_{i=1}^{n} p(y_i | z_i) \right] \mathcal{N}(\mathbf{z} | 0, \mathbf{K}) \, d\mathbf{z}$$



# Lower Bound optimization with SGD

$$\log \int \prod_{i=1}^{n} p(y_i|z_i)\mathcal{N}(\mathbf{z}|0,\mathbf{K})\ d\mathbf{z}$$

$$\mathcal{N}(\mathbf{z}|\mathbf{m},\mathbf{V})$$

$$\prod_{i=1}^{n} p(y_i|z_i)\mathcal{N}(\mathbf{z}|0,\mathbf{K})\ d\mathbf{z}$$

| | Natural parameters | Mean parameters |
|---|---|---|
| $\{\mathbf{m},\mathbf{V}\}$ | $\{\mathbf{V}^{-1}\mathbf{m}, -\frac{1}{2}\mathbf{V}^{-1}\}$ | $\{\mathbf{m}, \mathbf{V}+\mathbf{m}\mathbf{m}^T\}$ |

SGD's performance depends on parameterization.

Large number of parameters. Not modular.

# Conjugate-Computation VI

Converting the non-conjugate VI to a sequence of conjugate VI by using stochastic mirror-descent method

# CVI: Assumptions

- The posterior approximation is a minimal exponential family distribution

$$q(\mathbf{z}|\boldsymbol{\lambda}) := \exp\left\{\langle \boldsymbol{\phi}(\mathbf{z}), \boldsymbol{\lambda}\rangle - A(\boldsymbol{\lambda})\right\}$$

- Natural Parameter $\boldsymbol{\lambda}$
- Mean Parameter $\boldsymbol{\mu} := \mathbb{E}_q[\boldsymbol{\phi}(\mathbf{z})]$

# CVI : Main Ideas

SGD: $$\lambda_{t+1} = \lambda_t + \beta_t \; \frac{\partial \mathcal{L}_t}{\partial \lambda}$$

$$= \max_{\lambda} \left\langle \lambda, \frac{\partial \mathcal{L}_t}{\partial \lambda} \right\rangle - \frac{1}{\beta} \|\lambda - \lambda_t\|^2$$

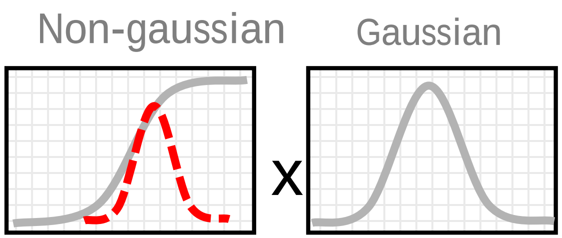
Optimize w.r.t. the mean parameter

Change the geometry to KL (natural gradient)

CVI: $$\mu_{t+1} = \max_{\mu} \left\langle \mu, \frac{\partial \mathcal{L}_t}{\partial \mu} \right\rangle - \frac{1}{\beta} \mathbb{D}_{KL}[q \| q_t]$$

# CVI gives simpler updates

$$p(\mathbf{z}|\mathbf{y}) \propto \prod_{i=1}^{n} p(y_i|z_i)\mathcal{N}(\mathbf{z}|0,\mathbf{K})$$




Compute gradient of the non-conjugate part

$$g_{1it} = \nabla_{m_i}\mathbb{E}_q[\log p(y_i|z_i)]$$

$$g_{2it} = \nabla_{m_i+V_{ii}}\mathbb{E}_q[\log p(y_i|z_i)]$$

$$q_{t+1}(\mathbf{z}) \propto \left[\prod_{i=1}^{n} e^{z_i g_{1it} + z_i^2 g_{2it}}\mathcal{N}(\mathbf{z}|0,\mathbf{K})\right]^{1-\beta_t} q_t(\mathbf{z})^{\beta_t}$$

New Posterior Distribution | Data | Prior | Previous Posterior Distribution

- No need to compute the gradient of the conjugate parts.
- Convert non-conjugate terms to conjugate terms

# Main features of CVI

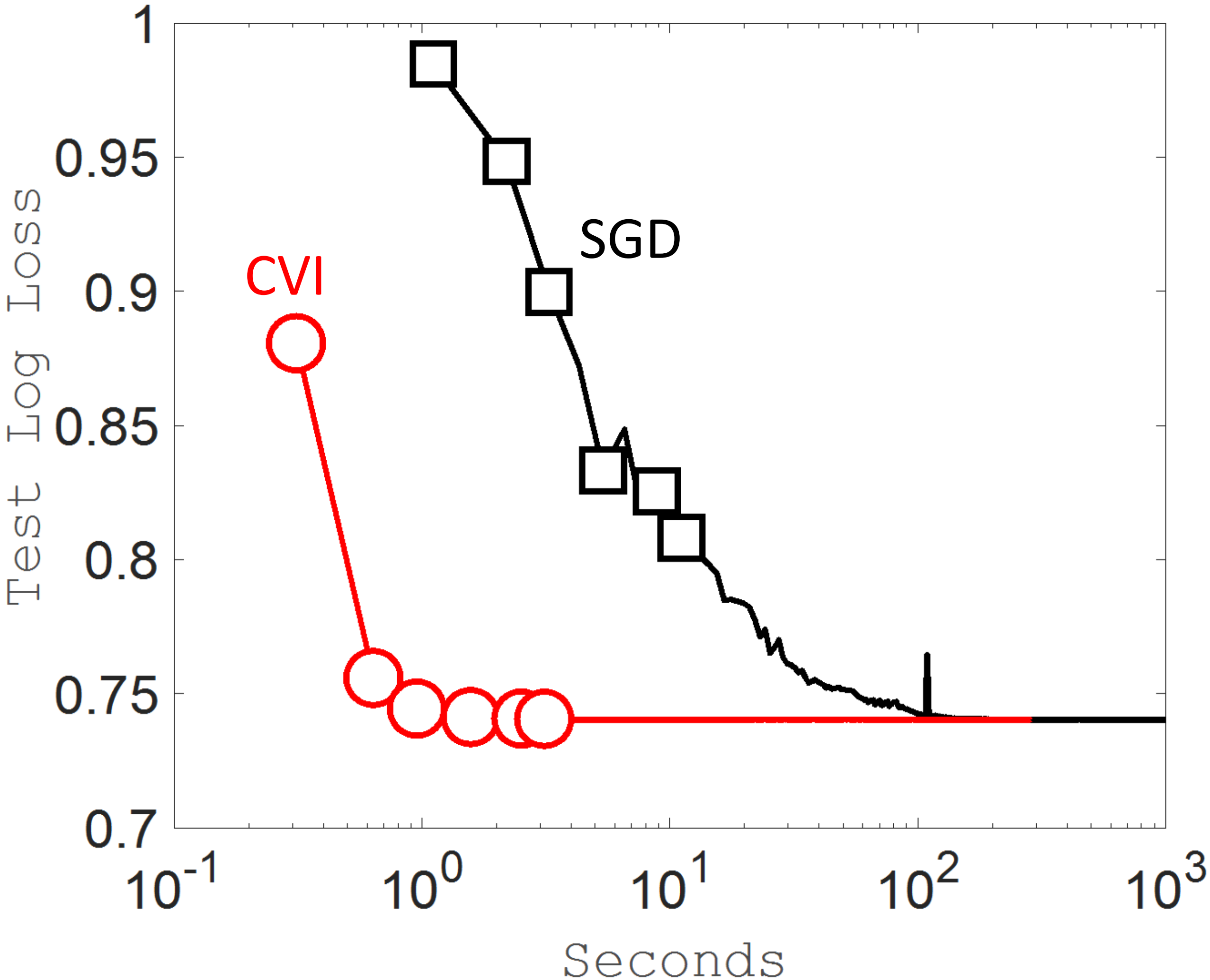
$$q_{t+1}(\mathbf{z}) \propto \left[\prod_{i=1}^{n} e^{z_i g_{1it} + z_i^2 g_{2it}} \mathcal{N}(\mathbf{z}|0, \mathbf{K})\right]^{1-\beta_t} q_t(\mathbf{z})^{\beta_t}$$

- Invariant to parameterization
- Express as a Bayesian model (comp. efficiency)
- For mean-field approximations, we can use message-passing (modularity)
- We can use “doubly” stochastic updates
- Enables structured inference in deep models!

# Related Work

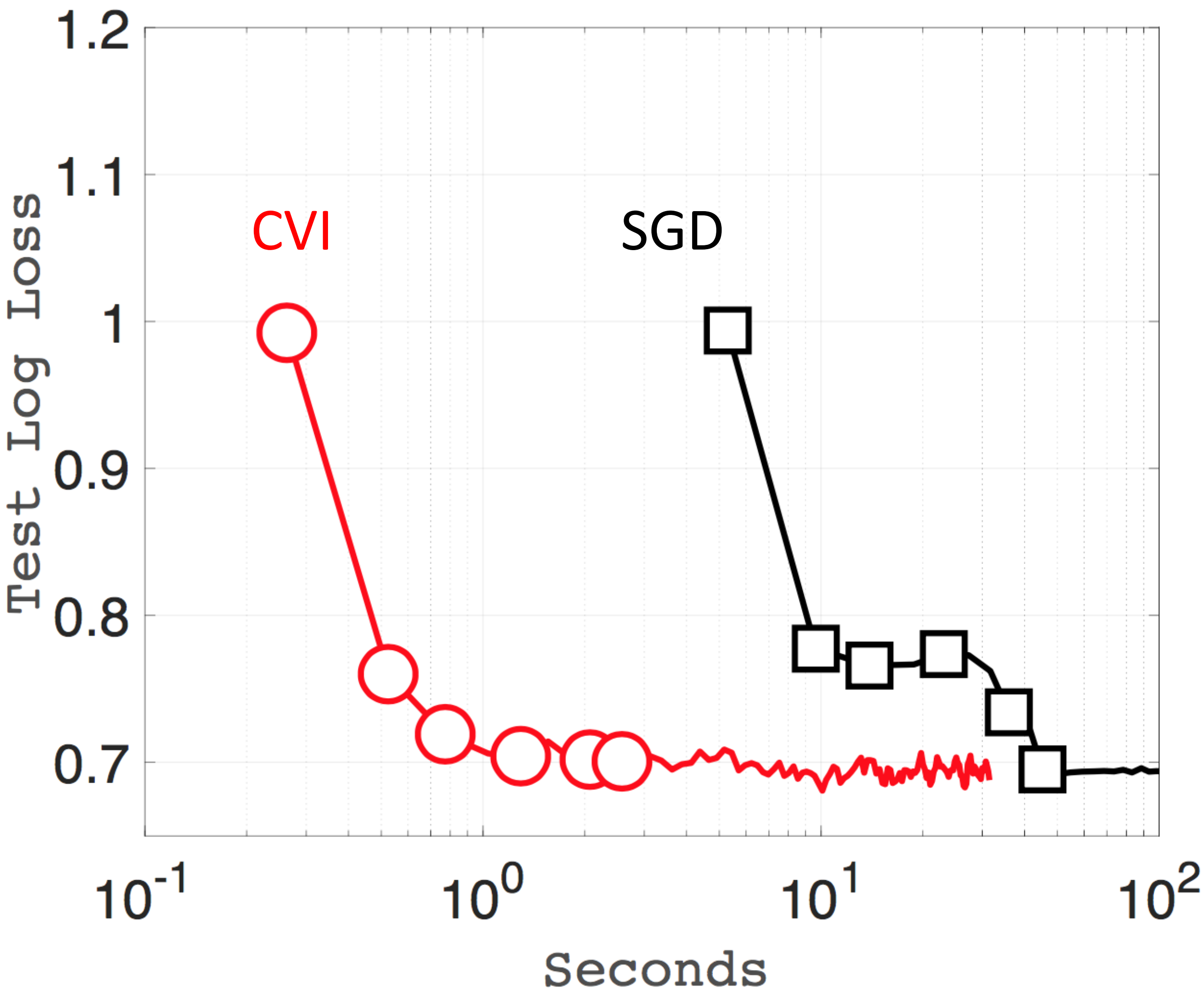
1. VMP (Winn et.al. 2005) and SVI (Hoffman et. al. 2013) do not apply to non-conjugate models.
2. Non-conjugate VMP (Minka et. al. 2011) does not allow stochastic gradient and lacks convergence guarantees.
3. EP (Minka 2001) has the same issues.
4. Naive SGD based methods do not always have easy to implement updates, e.g. Black-Box Variational Inference (BBVI) (Rangnathan et.al. 2014),
5. Salimans and Knowles 2014 is very similar, but require computation and storage of Fisher information matrix.

# Logistic Regression n>p



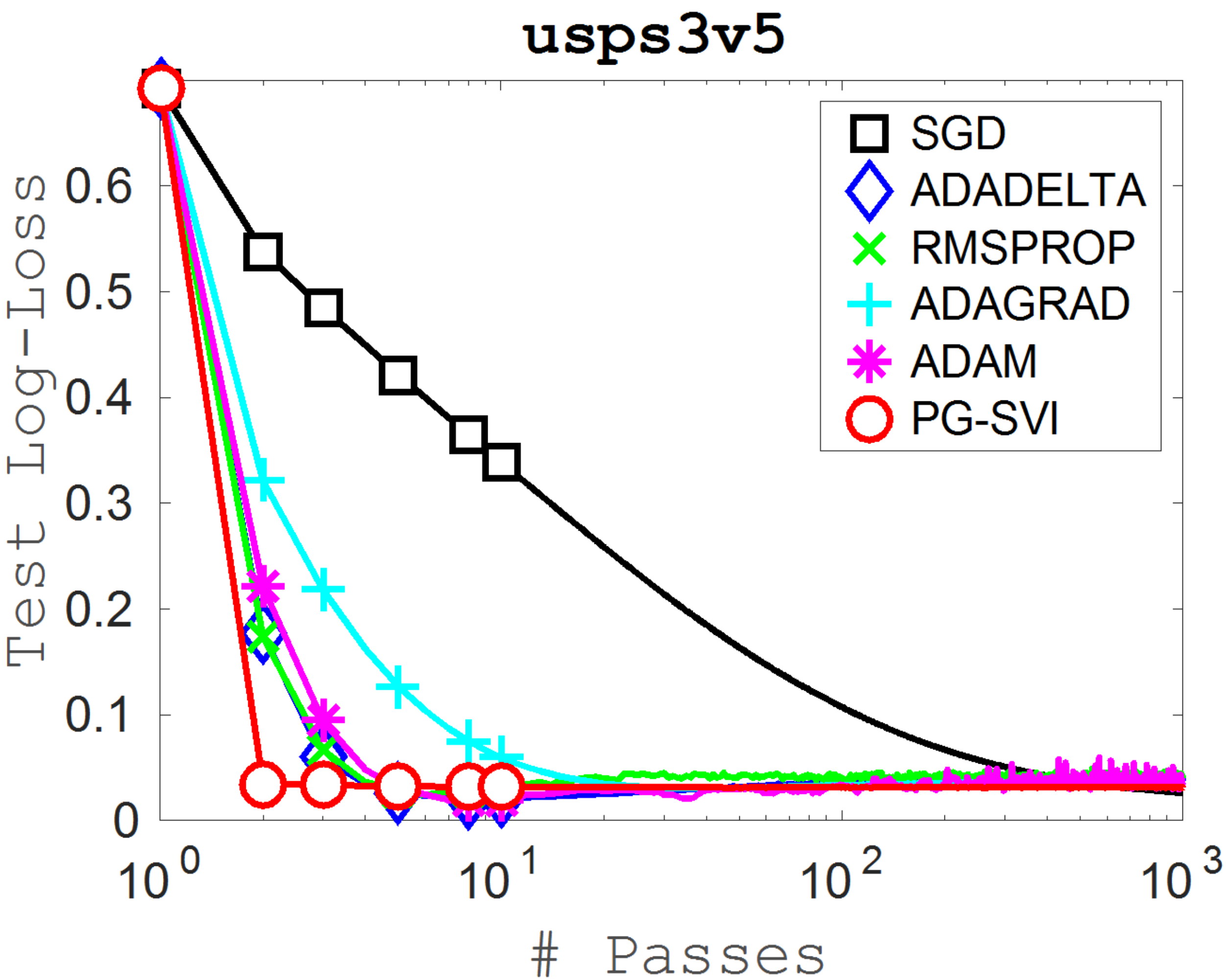

Covtype Scale dataset
n = 581,012
p = 54

# Logistic Regression n<p




Colon-cancer
n = 62
p = 2000

# Gaussian Process Classification




Gaussian process classification on 'USPS dataset' n = 1781

# Thanks for listening!

Code available at https://github.com/emtiyaz/cvi/

*Conjugate-Computation Variational Inference : Converting Variational Inference in Non-Conjugate Models to Inferences in Conjugate Models*, (AIstats 2017) M.E. Khan and Wu Lin

I am looking for post-docs, research scientist, and interns!

Visit my page at https://emtiyaz.github.io