Conjugate-Computation Variational Inference

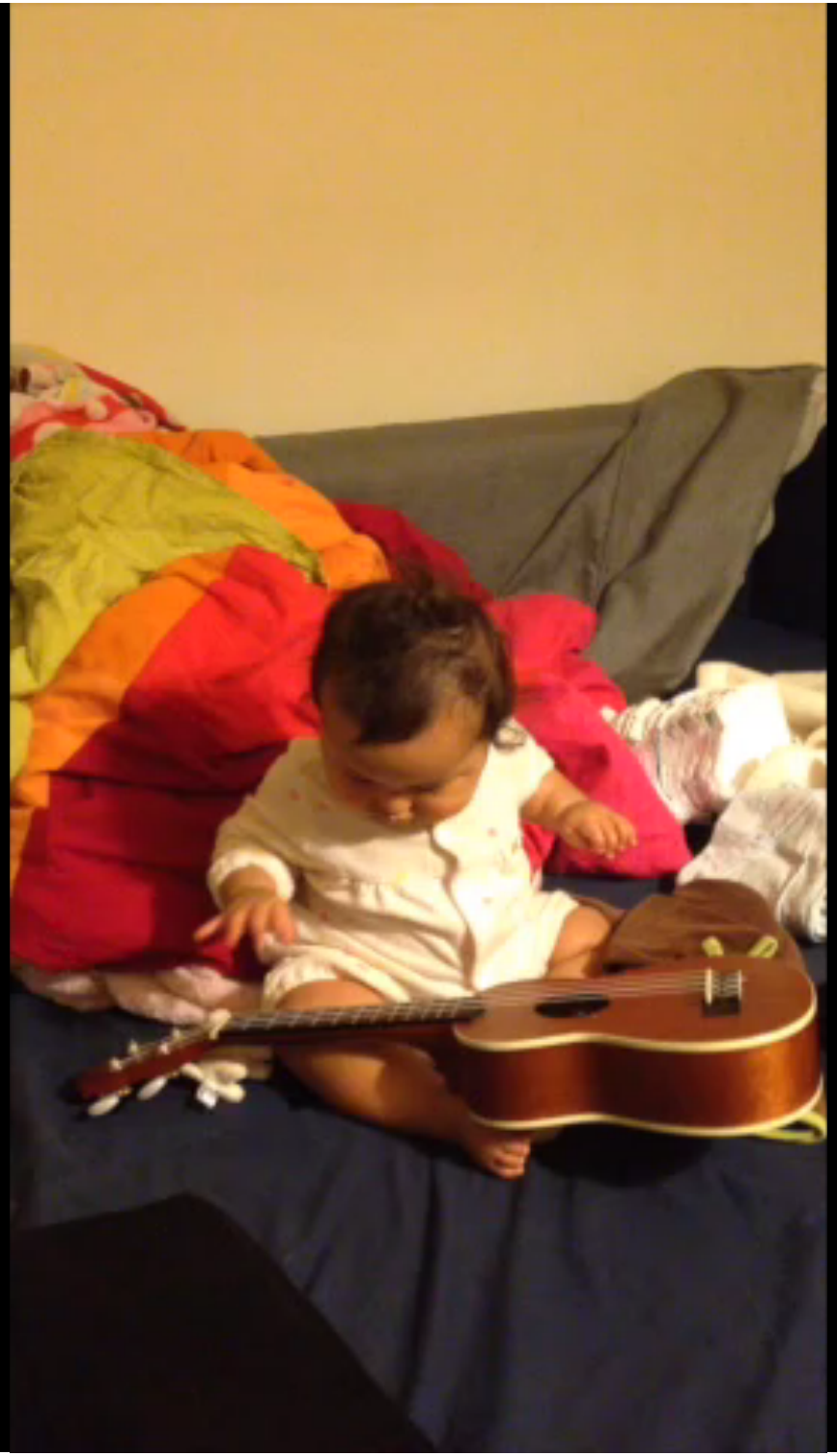
for approximate Bayesian inference in non-conjugate exponential-family models

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

Joint work with Wu Lin (to appear in AI-Stats 2017)
Find the pdf of this presentation at
https://emtiyaz.github.io

My research

"To understand the fundamental principles of learning from data and use them to develop algorithms that can learn like living beings."

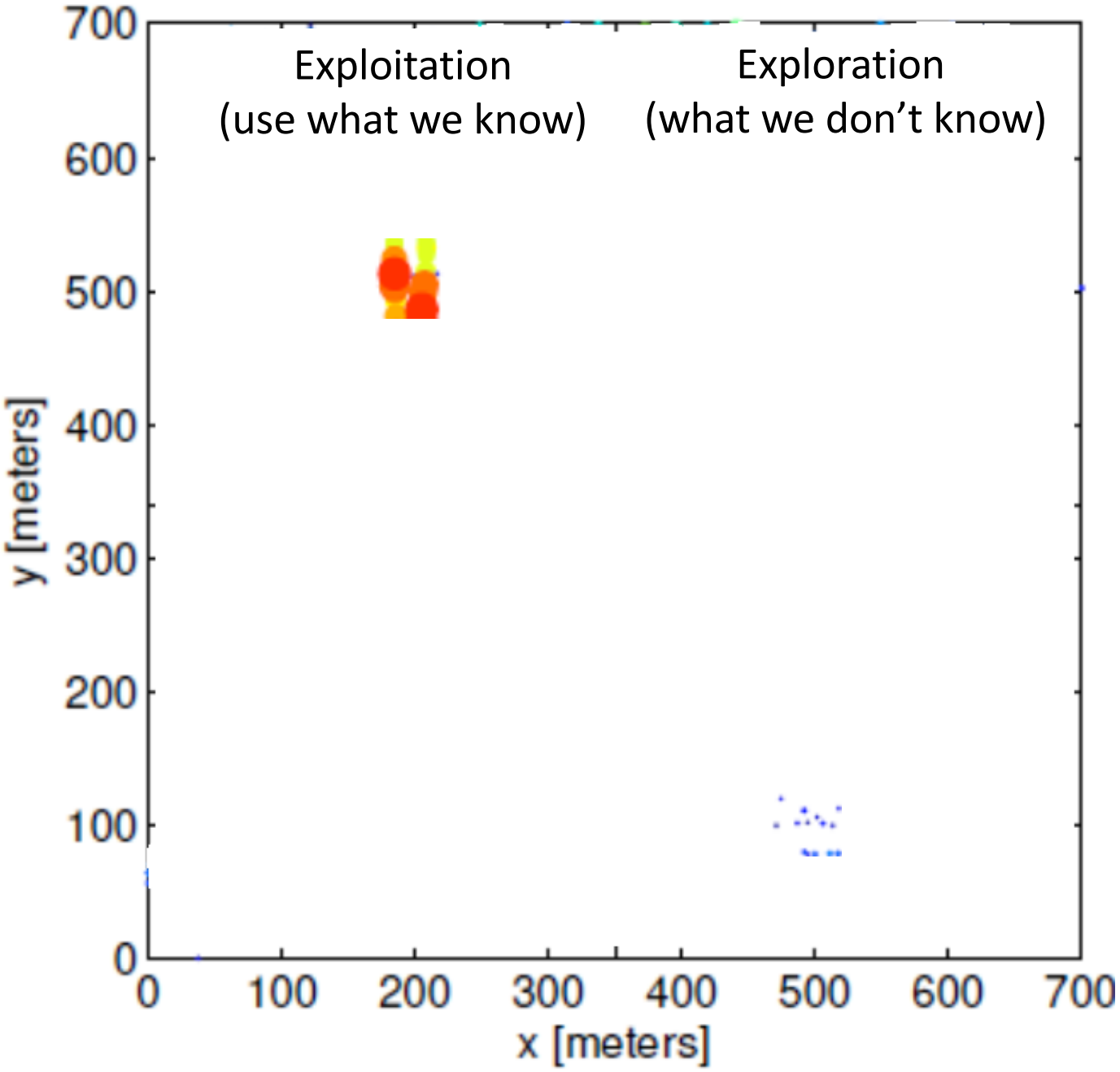


6 months old, learning about ukulele

Uncertainty

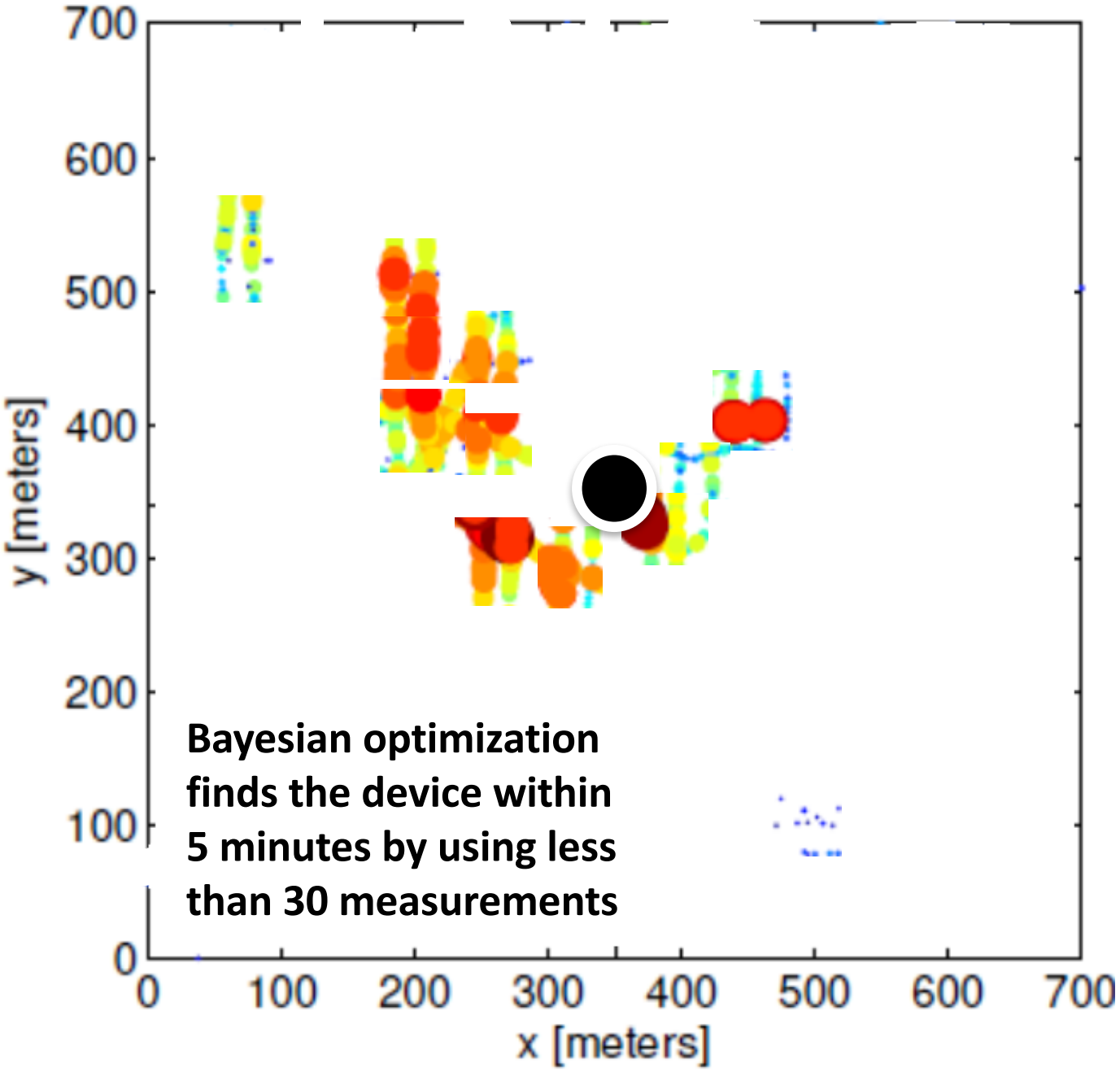
“How much we don’t know about the things we don’t know.”

Searching for a Cellphone (NIPS workshop 2015)



Bayesian Optimization

Searching for a Cellphone (NIPS workshop 2015)



Bayesian optimization finds the device within 5 minutes by using less than 30 measurements

Uncertainty is useful in exploration

Uncertainty Estimation is Computationally Challenging

Bayes' rule

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

Unknown Known

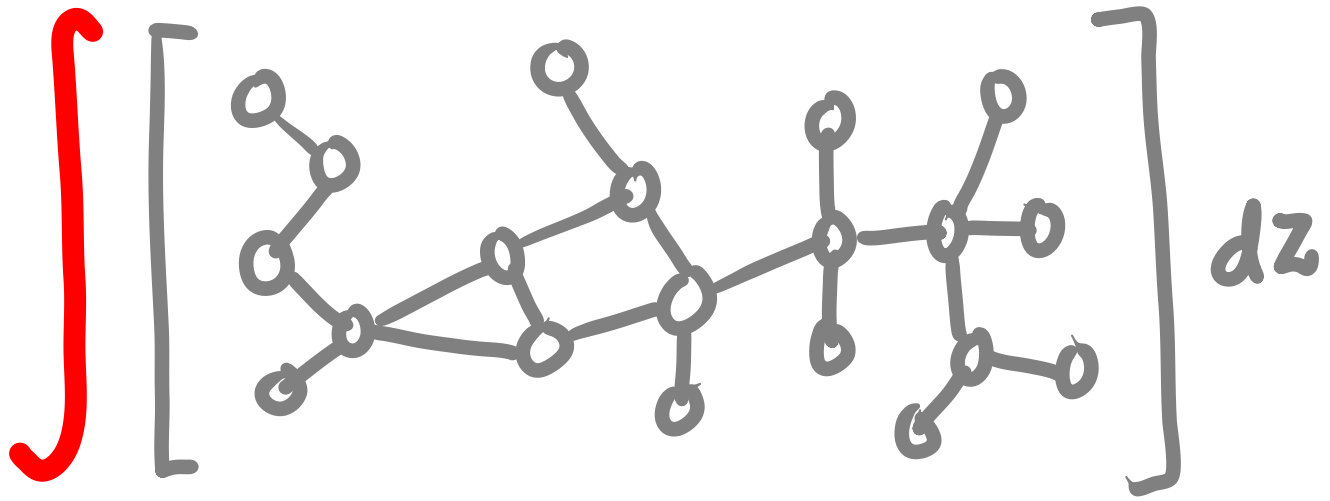
Exact computation of the integral is difficult.

Optimization

To approximate the integral

Approximate Bayesian Inference

Variational Inference: Integration to Optimization



$$\geq \max_{q} \int [\circ\circ] q\, dz_1 + \int [\circ\circ] q\, dz_2 + \cdots + \int [\circ\circ] q\, dz_3$$

High-dimensional “intractable” lower bound optimization

Stochastic Gradient Descent

General and scalable, but not always computationally efficient and modular, and sometimes dependent on parameterization.

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \beta_t \frac{\partial \mathcal{L}(\boldsymbol{\lambda}_t)}{\partial \boldsymbol{\lambda}}$$

Conjugate-Computation VI

- Scalable, general, efficient, modular, easy to implement, and convergent.
- By using proximal-gradient method, express the update as an "inference in a conjugate model".
  - Logistic Regression to Linear Regression
  - GP classification to GP Regression
  - Advanced Topic model to LDA
- In general, for variational inference in a general graphical model we can use message passing.

Outline

- Example of a non-conjugate model
- Challenges of variational inference (VI)
- Conjugate-Computation VI
  - Non-conjugate exp-family models
  - Extension to conditionally-conjugate models
  - Related work and convergence
  - Results
- Conclusions and future work.

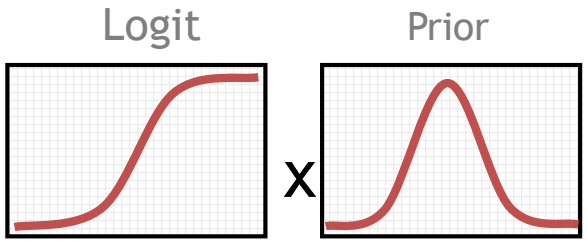
Non-conjugate models

Bayes' rule

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

Gaussian Proess classification (GPC)

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



Approximate Bayesian Inference

Variational inference simplifies the computation by restricting the posterior

$$p(\boldsymbol{z}|\boldsymbol{y}) \approx q(\boldsymbol{z})$$

VI: Integration to Optimization

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

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

$$\int \left[\log \frac{\mathcal{N}(\quad)}{q(z)} \right] q(z)\, dz$$

$$\max_{\boldsymbol{\lambda}} \quad \sum_{i=1}^{n} \mathbb{E}_q[\log p(y_i|z_i)] - \mathbb{D}_{KL}\left[q(\mathbf{z}|\boldsymbol{\lambda}) \| \mathcal{N}(\mathbf{z}|0,\mathbf{K})\right]$$

Three Challenges

$$\max_{\boldsymbol{\lambda}} \quad \sum_{i=1}^{n} \mathbb{E}_q[\log p(y_i|z_i)] - \mathbb{D}_{KL}\left[q(\mathbf{z}|\boldsymbol{\lambda})||\mathcal{N}(\mathbf{z}|0,\mathbf{K})\right]$$

Non-conjugate (difficult)

Conjugate (easy)

Intractable integrals (non-conjugacy)

Too many intractable integrals (large y)

Too many variational parameters (large z)

Stochastic Gradient Descent

General and scalable, but not always computationally efficient and modular, and sometimes dependent on parameterization.

$$\max_{\boldsymbol{\lambda}} \quad \sum_{i=1}^{n} \mathbb{E}_q[\log p(y_i|z_i)] - \mathbb{D}_{KL}\left[q(\mathbf{z}|\boldsymbol{\lambda})||\mathcal{N}(\mathbf{z}|0,\mathbf{K})\right]$$

Suppose $q(\mathbf{z}|\boldsymbol{\lambda}) = \mathcal{N}(\mathbf{z}|\mathbf{m},\mathbf{V})$ then $\lambda := \{\mathbf{V}^{-1}\mathbf{m}, -\frac{1}{2}\mathbf{V}^{-1}\}$

or mean parameters $\mu := \{\mathbf{m}, \mathbf{V} + \mathbf{m}\mathbf{m}^T\}$

$$\mathbf{V} \leftarrow \mathbf{V} + \alpha\left[\sum_{i=1}^{n} \frac{\partial}{\partial V}\mathbb{E}_q[\log p(y_i|z_i)] + \mathbf{K}^{-1} - \mathbf{V}^{-1}\right]$$

Conjugate-Computation VI

Converting the non-conjugate VI to a sequence of conjugate VI by using stochastic proximal-gradient method

Khan and Lin, AI-Stats, 2017

Khan, Babanezhad, Lin, Schmidt, Sugiyama, UAI, 2016

CVI: Assumptions

- The posterior approximation is a minimal exponential family distribution

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

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

- Natural Parameter $\boldsymbol{\lambda} := \{\mathbf{V}^{-1}\mathbf{m}, -\frac{1}{2}\mathbf{V}^{-1}\}$
- Sufficient Statistics $\boldsymbol{\phi}(\mathbf{z}) := \{\mathbf{z}, \mathbf{z}\mathbf{z}^T\}$
- Mean Parameter $\boldsymbol{\mu} := \mathbb{E}_q[\boldsymbol{\phi}(\mathbf{z})] = \{\mathbf{m}, \mathbf{V} + \mathbf{m}\mathbf{m}^T\}$

CVI : Main Ideas

Start with gradient descent

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \beta_t \frac{\partial \mathcal{L}(\boldsymbol{\lambda}_t)}{\partial \boldsymbol{\lambda}}$$

$$\boldsymbol{\lambda}_{t+1} = \max_{\lambda} \left\langle \boldsymbol{\lambda}, \frac{\partial \mathcal{L}(\boldsymbol{\lambda}_t)}{\partial \boldsymbol{\lambda}} \right\rangle - \frac{1}{2\beta_t} \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_t\|^2$$

$$\boldsymbol{\mu}_{t+1} = \max_{\mu} \left\langle \boldsymbol{\mu}, \frac{\partial \mathcal{L}(\boldsymbol{\mu}_t)}{\partial \boldsymbol{\mu}} \right\rangle - \frac{1}{\beta_t} \mathbb{D}_{KL} \left[q(\mathbf{z}|\boldsymbol{\lambda}) \| q(\mathbf{z}|\boldsymbol{\lambda}_t) \right]$$

Optimize w.r.t. the mean parameter

Change the geometry to KL

CVI gives simpler updates

$$\boldsymbol{\mu}_{t+1} = \max_{\mu} \left\langle \boldsymbol{\mu}, \frac{\partial \mathcal{L}(\boldsymbol{\mu}_t)}{\partial \boldsymbol{\mu}} \right\rangle - \frac{1}{\beta_t} \mathbb{D}_{KL} \left[q(\mathbf{z}|\boldsymbol{\lambda}) \| q(\mathbf{z}|\boldsymbol{\lambda}_t) \right]$$

$$\mathcal{L}(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \mathbb{E}_q[\log p(y_i|z_i)] - \mathbb{D}_{KL} \left[q(\mathbf{z}|\boldsymbol{\lambda}) \| \mathcal{N}(\mathbf{z}|0, \mathbf{K}) \right]$$

Non-conjugate (difficult) Conjugate (easy)

Gradient of conjugate term = conjugate term

Gradient of non-conjugate term = conjugate term

Gradient of Conjugate = Conjugate

$$\sum_{i=1}^{n} \mathbb{E}_q[\log p(y_i|z_i)] + \mathbb{E}_q\left[\log \mathcal{N}(\mathbf{z}|0,\mathbf{K})\right] - \mathbb{E}_q\left[\log q(\mathbf{z})\right]$$

Non-conjugate (difficult)

Conjugate (easy)

$$\left\langle \mu, \frac{\partial}{\partial \mu} \mathbb{E}_{q(z|\lambda_t)}\left[\log \mathcal{N}(\mathbf{z}|0,\mathbf{K})\right] \right\rangle = \mathbb{E}_q\left[\log \mathcal{N}(\mathbf{z}|0,\mathbf{K})\right] + \text{constant}$$

$$\begin{aligned}
&\left\langle \mu, \frac{\partial}{\partial \mu} \mathbb{E}_{q(z|\lambda)}\left[\langle \boldsymbol{\phi}(\mathbf{z}), \eta\rangle - A(\eta)\right] \right\rangle \quad = \left\langle \mu, \frac{\partial}{\partial \mu}\left[\langle \mu, \eta\rangle - A(\eta)\right] \right\rangle \\
&= \langle \mu, \eta\rangle \quad = \mathbb{E}_{q(z|\lambda)}\left[\langle \boldsymbol{\phi}(\mathbf{z}), \eta\rangle\right] \\
&= \mathbb{E}_{q(z|\lambda)}\left[\langle \boldsymbol{\phi}(\mathbf{z}), \eta\rangle - A(\eta)\right] + \text{constant}
\end{aligned}$$

$$\left\langle \mu, \frac{\partial}{\partial \mu} \mathbb{E}_{q(z|\lambda_t)}\left[\log q(\mathbf{z}|\boldsymbol{\lambda}_t)\right] \right\rangle = \mathbb{E}_q\left[\log q(\mathbf{z}|\boldsymbol{\lambda}_t)\right] + \text{constant}$$

Gradient of Non-Conjugate = Conjugate

$$\sum_{i=1}^{n} \mathbb{E}_q[\log p(y_i|z_i)] + \mathbb{E}_q\left[\log \mathcal{N}(\mathbf{z}|0,\mathbf{K})\right] - \mathbb{E}_q\left[\log q(\mathbf{z})\right]$$

Non-conjugate (difficult) Conjugate (easy)

$$\left\langle \mu, \frac{\partial}{\partial \mu} \mathbb{E}_{q(z|\lambda)}\left[\log p(y_i|z_i)\right] \right\rangle = \mathbb{E}_q\left(z_i g_i^{(1)} + z_i^2 g_i^{(2)}\right) + \text{constant}$$

Recall $\mu := \mathbb{E}_q[\phi(\mathbf{z})] = \{\mathbf{m}, \mathbf{V} + \mathbf{m}\mathbf{m}^T\}$

$$\{m_i, V_{ii} + m_i^2\}$$

$$\frac{\partial}{\partial m_i}\mathbb{E}_q[\log p(y_i|z_i)] = \hat{g}_i^{(2)}$$

$$\frac{\partial}{\partial (V_{ii} + m_i^2)}\mathbb{E}_q[\log p(y_i|z_i)] = \hat{g}_i^{(2)}$$

$$= m_i g_i^{(1)} + (V_{ii} + m_i^2) g_i^{(2)}$$

$$= \mathbb{E}_q\left(z_i g_i^{(1)} + z_i^2 g_i^{(2)}\right)$$

$$= \mathbb{E}_q\left[\log \mathcal{N}(\tilde{y}_i|z_i, \tilde{\sigma}_i^2)\right] + \text{constant}$$

$$\text{where } \tilde{y}_i = -\frac{g_i^{(1)}}{2g_i^{(2)}} \text{ and } \tilde{\sigma}_i^2 = 1/(2g_i^{(2)})$$

Putting it all together

$$\mathcal{L}(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \mathbb{E}_q[\log p(y_i|z_i)] + \mathbb{E}_q\left[\log \mathcal{N}(\mathbf{z}|0,\mathbf{K})\right] - \mathbb{E}_q\left[\log q(\mathbf{z})\right]$$

$$\left\langle \boldsymbol{\mu}, \frac{\partial \mathcal{L}(\boldsymbol{\mu}_t)}{\partial \boldsymbol{\mu}} \right\rangle = \sum_{i=1}^{n} \mathbb{E}_q\left[z_i g_{it}^{(1)} + z_i^2 g_{it}^{(2)}\right] + \mathbb{E}_q\left[\log \mathcal{N}(\mathbf{z}|0,\mathbf{K})\right] - \mathbb{E}_q\left[\log q(\mathbf{z})\right]$$

$$\boldsymbol{\mu}_{t+1} = \max_{\mu} \left\langle \boldsymbol{\mu}, \frac{\partial \mathcal{L}(\boldsymbol{\mu}_t)}{\partial \boldsymbol{\mu}} \right\rangle - \frac{1}{\beta_t} \mathbb{D}_{KL}\left[q(\mathbf{z}|\boldsymbol{\lambda}) \| q(\mathbf{z}|\boldsymbol{\lambda}_t)\right]$$

$$\boldsymbol{\mu}_{t+1} = \max_{\mu} \sum_{i=1}^{n} \mathbb{E}_q\left[z_i g_{it}^{(1)} + z_i^2 g_{it}^{(2)}\right] + \mathbb{E}_q\left[\log \mathcal{N}(\mathbf{z}|0,\mathbf{K})\right] - \mathbb{E}_q\left[\log q(\mathbf{z})\right] + \frac{1}{\beta_t}\mathbb{E}_q[\log q(\mathbf{z}|\boldsymbol{\lambda}_t)] - \frac{1}{\beta_t}\mathbb{E}_q[\log q(\mathbf{z})]$$

Closed form updates

$$\boldsymbol{\mu}_{t+1} = \max_{\mu} \sum_{i=1}^{n} \mathbb{E}_q \left[z_i g_{it}^{(1)} + z_i^2 g_{it}^{(2)} \right] + \mathbb{E}_q \left[\log \mathcal{N}(\mathbf{z}|0, \mathbf{K}) \right] - \mathbb{E}_q \left[\log q(\mathbf{z}) \right]$$

$$+ \frac{1}{\beta_t} \mathbb{E}_q [\log q(\mathbf{z}|\boldsymbol{\lambda}_t)] - \frac{1}{\beta_t} \mathbb{E}_q [\log q(\mathbf{z})]$$

$$\mathbb{E}_q \left[\log \frac{\prod_{i=1}^{n} e^{z_{it} g_{it}^{(1)} + z_i^2 g_{it}^{(2)}} \; \mathcal{N}(\mathbf{z}|0, \mathbf{K}) \; q(\mathbf{z}|\boldsymbol{\lambda}_t)^{1/\beta_t}}{q(\mathbf{z})^{1+1/\beta_t}} \right]$$

$$r_t = 1/(1+\beta_t)$$

Maximum occurs when

$$q(\mathbf{z}|\boldsymbol{\lambda}_{t+1}) \propto \left[\prod_{i=1}^{n} e^{z_{it} g_{it}^{(1)} + z_i^2 g_{it}^{(2)}} \; \mathcal{N}(\mathbf{z}|0, \mathbf{K}) \right]^{1-r_t} q(\mathbf{z}|\boldsymbol{\lambda}_t)^{r_t}$$

New Posterior Distribution

Data

Prior

Previous Posterior Distribution

Doubly Stochastic Approximation

$$q(\mathbf{z}|\boldsymbol{\lambda}_{t+1}) \propto \left[\prod_{i=1}^{n} e^{z_{it} g_{it}^{(1)} + z_i^2 g_{it}^{(2)}} \; \mathcal{N}(\mathbf{z}|0,\mathbf{K})\right]^{1-r_t} q(\mathbf{z}|\boldsymbol{\lambda}_t)^{r_t}$$

Randomly select an example *j* and compute stochastic gradient $\hat{g}_{j,t}^{(1)}$ and $\hat{g}_{j,t}^{(2)}$

$$\frac{\partial}{\partial m_j}\mathbb{E}_q[\log p(y_j|z_j)] \approx \hat{g}_{j,t}^{(1)} = \frac{n}{S}\sum_{s=1}^{S}\frac{\partial}{\partial m_j}\log p(y_j|m_j + u^{(s)})$$

where u(s) are samples from standard normal.

$$q(\mathbf{z}|\boldsymbol{\lambda}_{t+1}) \propto \left[e^{z_{j,t}\hat{g}_{j,t}^{(1)} + z_i^2 \hat{g}_{j,t}^{(2)}} \; \mathcal{N}(\mathbf{z}|0,\mathbf{K})\right]^{1-r_t} q(\mathbf{z}|\boldsymbol{\lambda}_t)^{r_t}$$

Expressing as a conjugate model

$$q(\mathbf{z}|\boldsymbol{\lambda}_{t+1}) \propto \left[e^{z_{j,t}\hat{g}^{(1)}_{j,t} + z_i^2 \hat{g}^{(2)}_{j,t}} \; \mathcal{N}(\mathbf{z}|0,\mathbf{K}) \right]^{1-r_t} q(\mathbf{z}|\boldsymbol{\lambda}_t)^{r_t}$$

$$\widetilde{\lambda}^{(1)}_{i,t} = r_t \widetilde{\lambda}^{(1)}_{i,t-1} + (1-r_t)\delta_{i=j}\hat{g}^{(1)}_{j,t}$$

$$\widetilde{\lambda}^{(2)}_{i,t} = r_t \widetilde{\lambda}^{(2)}_{i,t-1} + (1-r_t)\delta_{i=j}\hat{g}^{(2)}_{j,t}$$

$$q(\mathbf{z}|\boldsymbol{\lambda}_0) := \mathcal{N}(\mathbf{z}|0,\mathbf{K})$$

Site parameter

$$q(\mathbf{z}|\boldsymbol{\lambda}_{t+1}) \propto \prod_{i \in I_t} e^{z_{j,t}\widetilde{\lambda}^{(1)}_{j,t} + z_i^2 \widetilde{\lambda}^{(2)}_{j,t}} \; \mathcal{N}(\mathbf{z}|0,\mathbf{K})$$

where *I_t* is the set of examples selected in the past

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

CVI for non-conjugate models

GP Classification:

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

$$q(\mathbf{z}|\boldsymbol{\lambda}_0) := \mathcal{N}(\mathbf{z}|0,\mathbf{K})$$

$$\widetilde{\lambda}^{(1)}_{i,t} = r_t \widetilde{\lambda}^{(1)}_{i,t-1} + (1-r_t)\delta_{i=j}\hat{g}^{(1)}_{j,t}$$
$$\widetilde{\lambda}^{(2)}_{i,t} = r_t \widetilde{\lambda}^{(2)}_{i,t-1} + (1-r_t)\delta_{i=j}\hat{g}^{(2)}_{j,t}$$

GP Regression:

$$q(\mathbf{z}|\boldsymbol{\lambda}_{t+1}) \propto \prod_{i\in I_t} e^{z_{j,t}\widetilde{\lambda}^{(1)}_{j,t} + z_i^2 \widetilde{\lambda}^{(2)}_{j,t}} \mathcal{N}(\mathbf{z}|0,\mathbf{K})$$

$$p(\mathbf{y},\mathbf{z}) \propto \tilde{p}_{nc}(\mathbf{y},\mathbf{z})\ \tilde{p}_c(\mathbf{y},\mathbf{z})$$

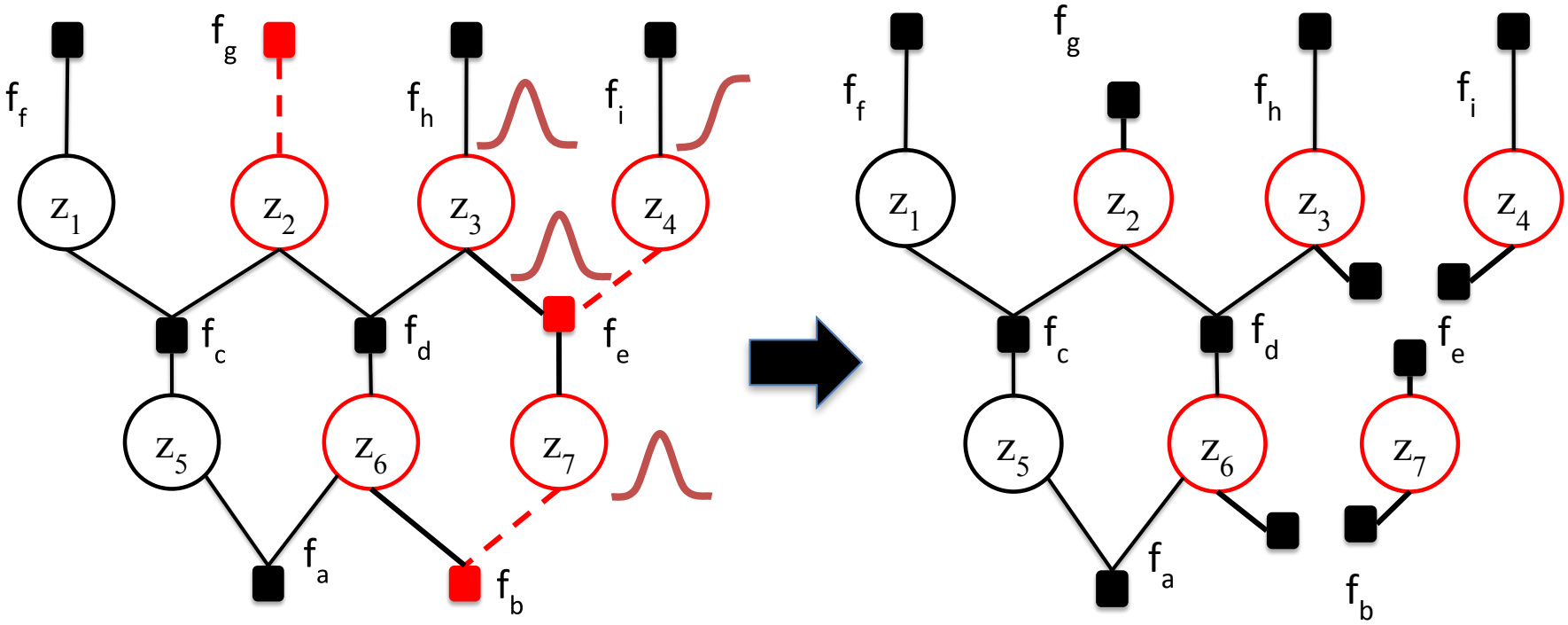
$$q(\mathbf{z}|\boldsymbol{\lambda}_0) \propto \tilde{p}_c(\mathbf{y},\mathbf{z})$$

$$\widetilde{\boldsymbol{\lambda}}_t = r_t \widetilde{\boldsymbol{\lambda}}_{t-1} + (1-r_t)\ \widehat{\nabla}_{\mu}\mathbb{E}_{q_t}\left[\log \tilde{p}_{nc}(\mathbf{y},\mathbf{z})\right]$$

$$q(\mathbf{z}|\boldsymbol{\lambda}_{t+1}) \propto \exp\left(\langle \boldsymbol{\phi}(\mathbf{z}), \widetilde{\boldsymbol{\lambda}}_t \rangle\right) \tilde{p}_c(\mathbf{y},\mathbf{z})$$

Extension to mean-field inference

When the model contains conditionally-conjugate factors



Assuming mean-field $q(\mathbf{z}) = \prod_k q(z_k)$, compute approximation:

$$\widetilde{\lambda}_{f_e \to z_k} \leftarrow r_t \widetilde{\lambda}_{f_e \to z_k} + (1 - r_t) \widehat{\nabla}_{\mu_k} \mathbb{E}_{q_t}[\log f_e]$$

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.
6. However, our method does not work for auto-encode type models yet (e.g. with recognition model).

Convergence

$$\mathbb{E}_{R,\xi}\left[\|(\boldsymbol{\lambda}_R - \boldsymbol{\lambda}_{R+1})/\beta\|^2\right] \leq \left[\frac{2LC_0}{\alpha_*^2 t} + \frac{c\sigma^2}{M\alpha_*}\right]$$

Faster Stochastic Variational Inference using Proximal-Gradient Methods with General Divergence Functions, (UAI 2016) M.E. Khan, R. Babanezhad, W. Lin, M. Schmidt, M. Sugiyama.

Based on Ghadimi, Lan and Zhang (2014)

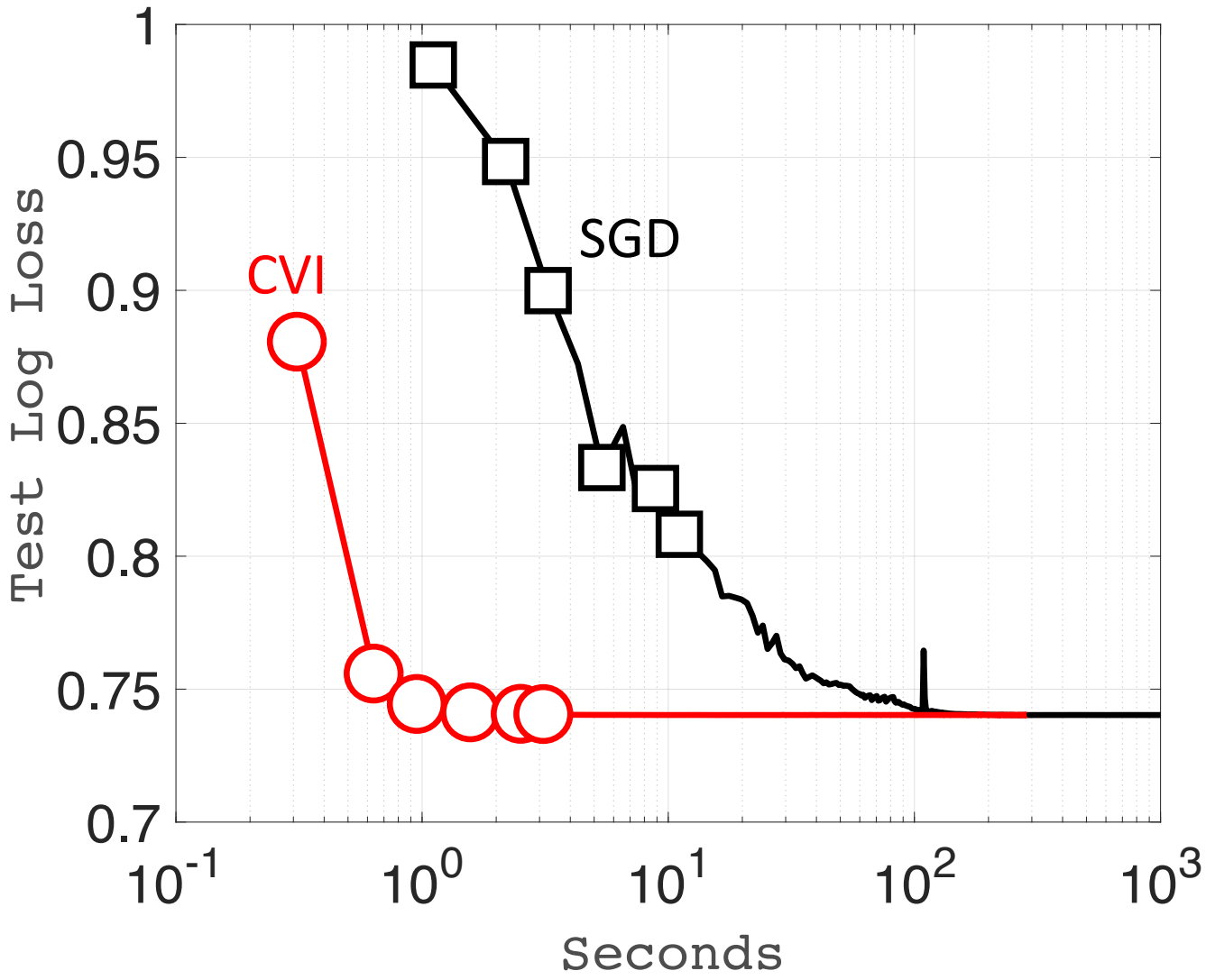
Gradient wrt the mean parameter

Based on the methods discussed in Salimans and Knowles (2014), we can compute the gradient w.r.t. the mean parameter.

1. Compute a stochastic approximation to the Fisher-information matrix.
2. Compute the gradient w.r.t. the natural parameter.
3. Solve the system (no need to store or form the matrix).

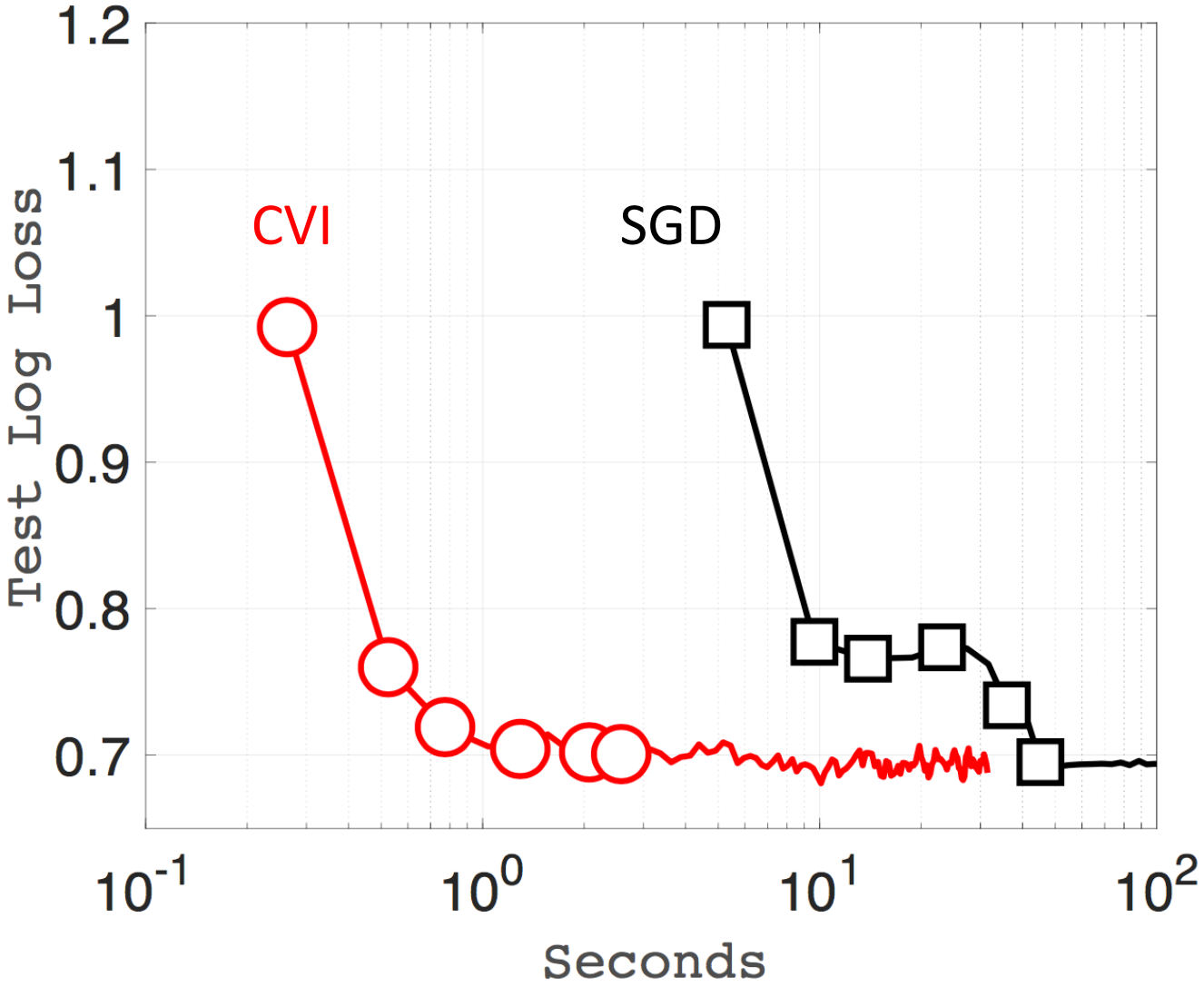
$$\nabla_\mu f(\mu) = \mathbf{I}(\boldsymbol{\lambda})^{-1} \nabla_\lambda f(\boldsymbol{\lambda})$$

Logistic Regression with large n



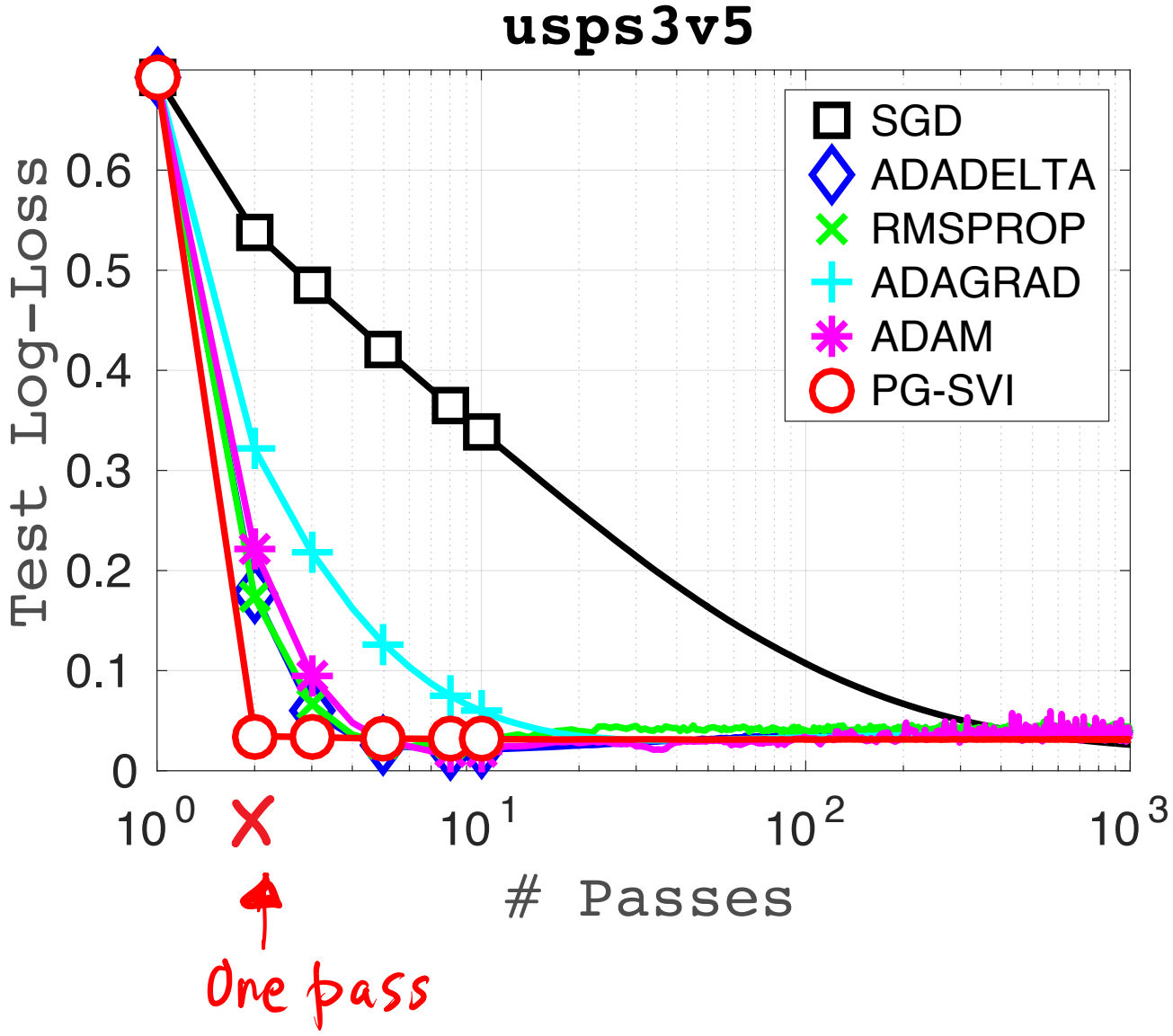
Covtype Scale dataset
n = 581,012
p = 54

Logistic Regression with large p



Colon-cancer
n = 62
p = 2000

Gaussian Process Classification



Gaussian process classification on 'USPS dataset' n = 1781

Conjugate-Computation VI

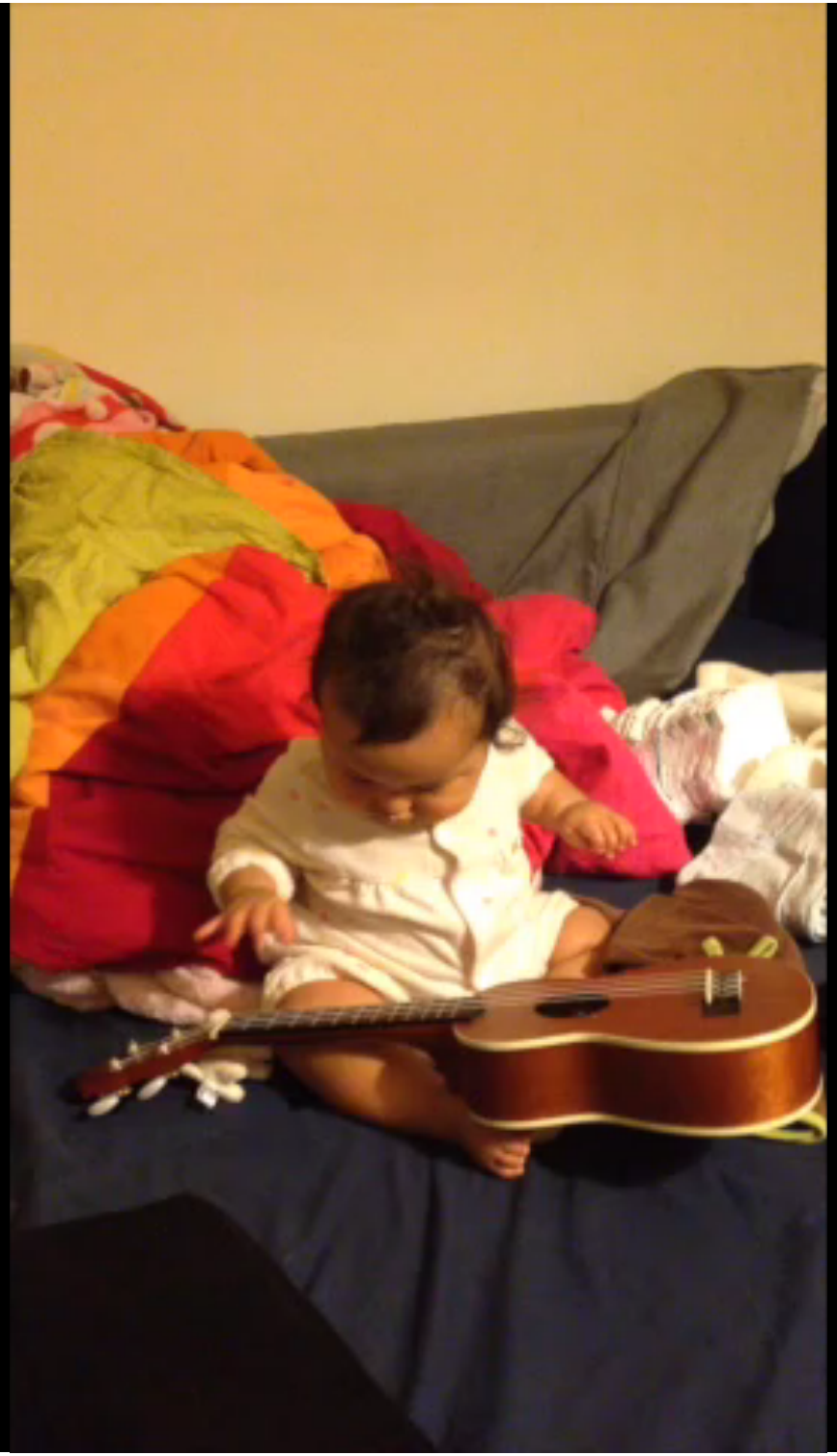
- Scalable, general, efficient, modular, easy to implement, and convergent.
- By using proximal-gradient method, express the update as an "inference in a conjugate model".
  - Logistic Regression to Linear Regression
  - GP classification to GP Regression
  - Advanced Topic model to LDA
- In general, for variational inference in a general graphical model we can use message passing.

On-going work

- Extension to structured mean-field.
- Application to deep neural network.
- Large-scale inference on GP models (and time-series models).
- Application to deep GP.
- Implementation for large and complex models.

Uncertainty

“How much we don’t know about the things we don’t know.”



6 months old, learning about ukulele

At the age of
12 months

Transfer
Learning
at 14 months

To Discover fundamental principles of learning from data

Bridging the gap between
“algorithms” and “babies”.

Some open questions

- How to collect relevant, good-quality data?
- How to sequentially and reliably learn from streaming but redundant data?
- How to deal with noisy, unreliable data?
- How to communicate with the environment to get feedback?

Thanks for listening!

https://emtiyaz.github.io

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

Faster Stochastic Variational Inference using Proximal-Gradient Methods with General Divergence Functions, (UAI 2016) M.E. Khan, R. Babanezhad, W. Lin, M. Schmidt, M. Sugiyama.

I am looking for post-docs/ research scientists!