A Fixed-Point Operator for Inference in Variational Bayesian Latent Gaussian Models: Supplementary Material

Rishit Sheth
rishit.sheth@tufts.edu

Roni Khardon
roni@cs.tufts.edu

Department of Computer Science, Tufts University, Medford, MA, USA

1 Datasets

The datasets used in the experiments are described in Table 1. We selected some medium size datasets from the literature to start with and added large ones to demonstrate performance in GLM. The samples and number of features columns specified in the table refer to the maximum sizes used in the experiments. Categorical features were converted using dummy coding. All features in all datasets were normalized using Z-scores. For regression, the target values were also Z-score normalized.

When both training and test sets were available, only data in the training sets were used except where noted in the following. The *epsilon* dataset used in these experiments was constructed from the first 15,000 samples and 1000 features of the original *epsilon* test set. The *ucsdpeds* dataset refers to the *ucsdpeds1l* dataset in the nomenclature of Chan & Vasconcelos (2012). We generated artificial count labels for the *epsilon-count* dataset as follows. We used the training data of *epsilon* and a GLM with Poisson likelihood and log link function. The parameter was sampled from the prior described in the experimental section. The dataset *cahousing* is referred to as *cadata* on `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`. The *yearpred* dataset consisted of the unique rows of the original test set. The *wlan-long* dataset was derived from the *UJIIndoorLoc* dataset from the UCI Machine Learning Repository by selecting unique rows and using longitude as the target variable. The *wlan-inout* dataset was derived similarly but by selecting inside vs. outside as the binary target variable.

2 Natural Gradients for LGM

This section shows that "FP-like" updates arise from natural gradients whenever the KL term is taken over distributions in the same exponential family. Similar derivations exist in the literature, so the analysis is not new. But here we emphasize the fixed point aspect of the update. We then derive the concrete natural gradient updates for LGM showing that the update for V is identical to the FP update in the main paper (as pointed out by an anonymous reviewer) and showing the corresponding "FP-like" update for $\boldsymbol{m}$. Please see discussion in the main paper for context and further details.

2.1 The General Form

The VLB for the LGM model is given by

$$\text{VLB} = \sum_i E_{q_i(f_i)}(\log \phi_i(f_i)) - \text{KL}(q(w)\|p(w)) \quad (1)$$

where in our main derivation $q_i(f_i) = \mathcal{N}(f_i|m_i, v_i)$, $m_i = a_i + d_i^T m$ and $v_i = c_i + d_i^T V d_i$, $q(w) = \mathcal{N}(w|m, V)$, and $p(w) = \mathcal{N}(w|\mu, \Sigma)$.

More generally, for distributions $p(w)$ and $q(w)$ of the same exponential family type,

$$p(w) = \exp\left(t(w)^T\theta_p - F(\theta_p)\right) h(w) \quad (2)$$

$$q(w) = \exp\left(t(w)^T\theta_q - F(\theta_q)\right) h(w) \quad (3)$$

the Kullback-Liebler divergence between $q(w)$ and $p(w)$ is given by

$$\text{KL}(q\|p) = \eta_q^T(\theta_q - \theta_p) - (F(\theta_q) - F(\theta_p)) \quad (4)$$

where η_q denotes the expectation (mean) parameters of q, i.e., $E_q(t(w))$.

The natural gradient update of the canonical (natural) parameters for $q(w)$ is given by

$$\theta_q \leftarrow \theta_q + I(\theta_q)^{-1}\frac{\partial \text{VLB}}{\partial \theta_q} \quad (5)$$

$$= \theta_q + I(\theta_q)^{-1}\frac{\partial \eta_q}{\partial \theta_q}\frac{\partial \text{VLB}}{\partial \eta_q} \quad (6)$$

$$= \theta_q + \frac{\partial \text{VLB}}{\partial \eta_q} \quad (7)$$

Table 1: Summary of data sets

Name	Samples	Features	Model Type	Source
a9a	32561	123	Binary	Lichman (2013)
epsilon	15000	1000	Binary	Sonnenburg et al. (2008)
madelon	2600	500	Binary	Guyon et al. (2004)
musk	6598	166	Binary	Lichman (2013)
usps (3s v. 5s)	1540	256	Binary	Rasmussen & Nickisch (2013)
wlan-inout	19085	466	Binary	Lichman (2013)
abalone	4177	8	Count	Lichman (2013)
flares	1065	24	Count	Lichman (2013)
ucsdpeds	4000	30	Count	Chan & Vasconcelos (2012)
cahousing	20640	8	Regression	Statlib (2015)
cpusmall	8192	12	Regression	Statlib (2015)
spacega	3107	6	Regression	Statlib (2015)
wlan-long	19085	466	Regression	Lichman (2013)
yearpred	51609	90	Regression	Lichman (2013)

since $\frac{\partial \eta}{\partial \theta} = I(\theta)$ for dual coordinate systems, θ and η (Amari & Nagaoka, 2000).

The derivative of the KL divergence with respect to the expectation parameters is given by

$$\frac{\partial \mathrm{KL}(q\|p)}{\partial \eta_q} = \theta_q - \theta_p + \left(\frac{\partial \theta_q}{\partial \eta_q}\right)^T \eta_q - \left(\frac{\partial \theta_q}{\partial \eta_q}\right)^T \frac{\partial F(\theta_q)}{\partial \theta_q} \tag{8}$$

$$= \theta_q - \theta_p \tag{9}$$

since $\frac{\partial F(\theta_q)}{\partial \theta_q} = \eta_q$ in the exponential family.

Now, denoting $A(\eta_q) = \frac{\partial}{\partial \eta_q}[\sum_i E_{q_i(f_i)}(\log \phi_i(f_i))]$ where we have emphasized the dependence on η_q, and applying this notation to (7) we get the "FP-like" update

$$\theta_q \leftarrow \theta_q - [\theta_q - \theta_p] + A(\eta_q) = \theta_p + A(\eta_q) \tag{10}$$

2.2 Natural Gradients for LGM

Recall that for the Gaussian distribution we have $\theta = (\mathbf{r}, S) = (V^{-1}\boldsymbol{m}, \frac{1}{2}V^{-1})$ and $\eta = (h, H) = (\boldsymbol{m}, -(V + \boldsymbol{m}\boldsymbol{m}^T))$. To take the derivative of the sum of expectations term in Eq. 1, we rewrite m_i and v_i with respect to the expectation parameters η_q

$$m_i = a_i + d_i^T h \tag{11}$$

$$v_i = c_i - d_i^T (H + hh^T) d_i \tag{12}$$

Note that v_i now depends on both expectation parameters whereas in the original (source) parameterization v_i only depended on one parameter, V.

The derivatives of the sum of expectations term are now given by

$$\begin{aligned} & \frac{\partial}{\partial h}\left[\sum_i E_{\mathcal{N}(f_i|m_i,v_i)}(\log \phi_i(f_i))\right] \\ = & \left(\frac{\partial m_i}{\partial h}\frac{\partial}{\partial m_i} + \frac{\partial v_i}{\partial h}\frac{\partial}{\partial v_i}\right)\left[\sum_i E_{\mathcal{N}(f_i|m_i,v_i)}(\log \phi_i(f_i))\right] \\ = & \sum_i (\rho_i + (h^T d_i)\gamma_i) d_i \end{aligned} \tag{13}$$

and

$$\begin{aligned} & \frac{\partial}{\partial H}\left[\sum_i E_{\mathcal{N}(f_i|m_i,v_i)}(\log \phi_i(f_i))\right] \\ = & \frac{\partial v_i}{\partial H}\frac{\partial}{\partial v_i}\left[\sum_i E_{\mathcal{N}(f_i|m_i,v_i)}(\log \phi_i(f_i))\right] \\ = & \sum_i \frac{1}{2}\gamma_i d_i d_i^T \end{aligned} \tag{14}$$

with

$$\rho_i = \frac{\partial}{\partial m_i} E_{\mathcal{N}(f_i|m_i,v_i)}(\log \phi_i(f_i)) \tag{15}$$

$$\gamma_i = -2\frac{\partial}{\partial v_i} E_{\mathcal{N}(f_i|m_i,v_i)}(\log \phi_i(f_i)) \tag{16}$$

Finally, the updates described by Eq. 7 are

$$V^{-1}\boldsymbol{m} \leftarrow \Sigma^{-1}\mu + \sum_i (\rho_i + (\boldsymbol{m}^T d_i)\gamma_i) d_i \tag{17}$$

$$\frac{1}{2}V^{-1} \leftarrow \frac{1}{2}\Sigma^{-1} + \sum_i \frac{1}{2}\gamma_i d_i d_i^T \tag{18}$$

Now (18) is identical to the FP update in the main paper, whereas (17) is a size 1 natural gradient step for $\boldsymbol{m}$. As discussed in the main paper, while (18) is analyzed and shown to work well empirically, our exploratory experiments with (17) showed that it does

not always converge to the optimal point. Experimental evidence illustrating this point is shown in the next section. Hence, in contrast with our analysis of FP for V, size 1 natural gradient updates do not provide a full explanation to the success of FP.

3 Additional Experimental Results

This section includes additional experimental results that were omitted from the main paper due to space constraints.

For all experiments in the main paper and supplementary material, the stopping conditions are $\|\nabla f(x_k)\|_\infty \leq 10^{-5}$, $f(x_{k-1}) - f(x_k) \leq 10^{-9}$, or $k > 500$ where f is the objective function being optimized, k represents the iteration number, and x is the current optimization variable.

Figure 1 shows monotonicity maps for additional likelihood functions demonstrating the same pattern: large continuous regions of the (m, v) space with the same direction of change. This shows that small changes to (m, v) are likely to be stable with respect to condition 2.

Figure 2 shows evidence of FP cycling in a GLM with the logistic likelihood trained on *wlan-inout*. Here, $\mathbf{m}$ was fixed to the optimal $\mathbf{m}^\star$ and V was initialized to I. Plots for other randomly selected γs are similar.

Figure 3 shows results for an incremental optimization with FP for both the covariance and the mean. We see that this method sometimes converges to the optimum, sometimes converges to an inferior point, and sometimes diverges (in the left-most plot, the VLB for this method increases out of the y-axis range). In contrast, FPb and FPi appear to be stable across the range of experimental conditions, and the same holds for GRAD although it is generally slower.

Figure 4 shows results for GLM on datasets which are smaller than the ones in the main paper. In this case, the performance of FP and GRAD is is not dramatically different. However, for the larger dataset in the main paper, FPi converges much faster.

To further explore performance on larger datasets, we have run multiple experiments with the *epsilon* dataset, where a subset of the features was randomly selected. The results for several such settings are shown in Figure 5. As can be seen, the difference between the algorithms becomes more pronounced when the number of features increases in this manner.

References

Amari, S. and Nagaoka, H. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical monographs*. Oxford University Press, 2000.

Chan, A. B. and Vasconcelos, N. Counting People With Low-Level Features and Bayesian Regression. *IEEE Transactions on Image Processing*, 21 (4):2160–2177, April 2012.

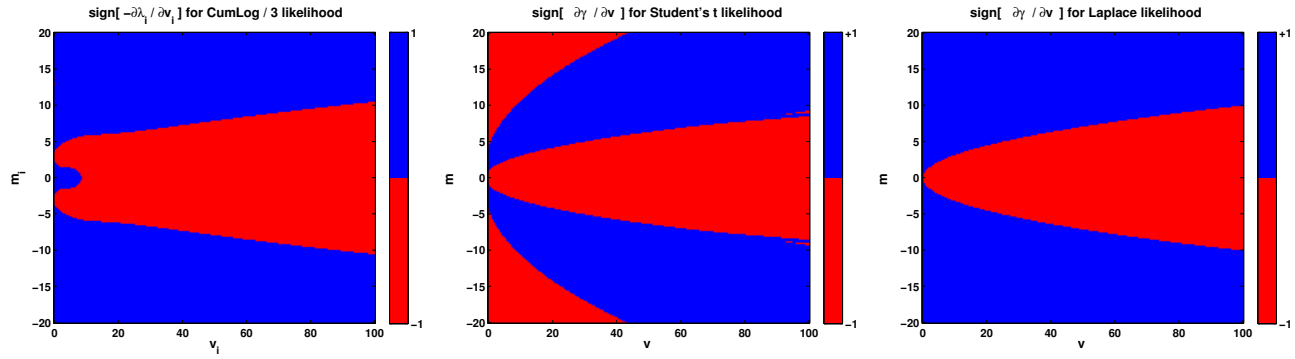
Guyon, Isabelle, Gunn, Steve, Ben-Hur, Asa, and Dror, Gideon. Result analysis of the NIPS 2003 feature selection challenge. In *Advances in neural information processing systems*, pp. 545–552, 2004.

Lichman, M. UCI machine learning repository, 2013. `http://archive.ics.uci.edu/ml`.

Rasmussen, Carl Edward and Nickisch, Hannes. GPML software package, 2013. `http://www.gaussianprocess.org/gpml/code/matlab/doc/`.

Sonnenburg, Soeren, Franc, Vojtech, Yom-Tov, Elad, and Sebag, Michele. Pascal large scale learning challenge, ICML Workshop, 2008. `http://largescale.first.fraunhofer.de`.

Statlib. Statlib datasets, 2015. Downloaded from `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`.



(a) (b) (c)

Figure 1: Plots of sign$(\frac{\partial}{\partial v}\gamma(v))$ for several observation likelihoods.

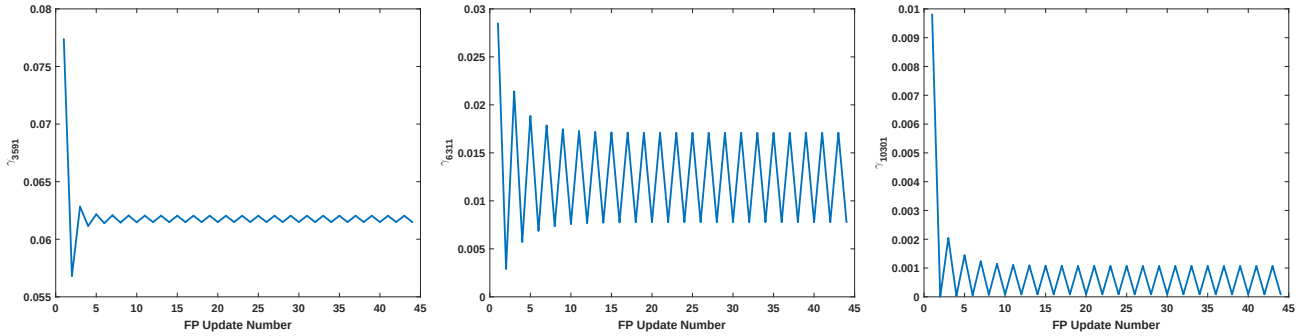


Figure 2: Plot of γ_i for $i = 3591, 6311$, and 10301 (out of 19085) vs. FP update number.

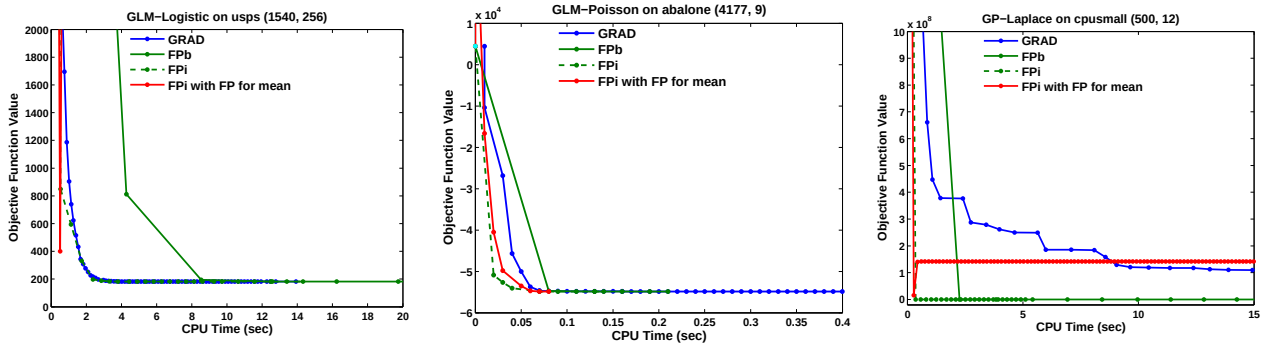


Figure 3: Comparison of using FP for the mean against other methods for three datasets.

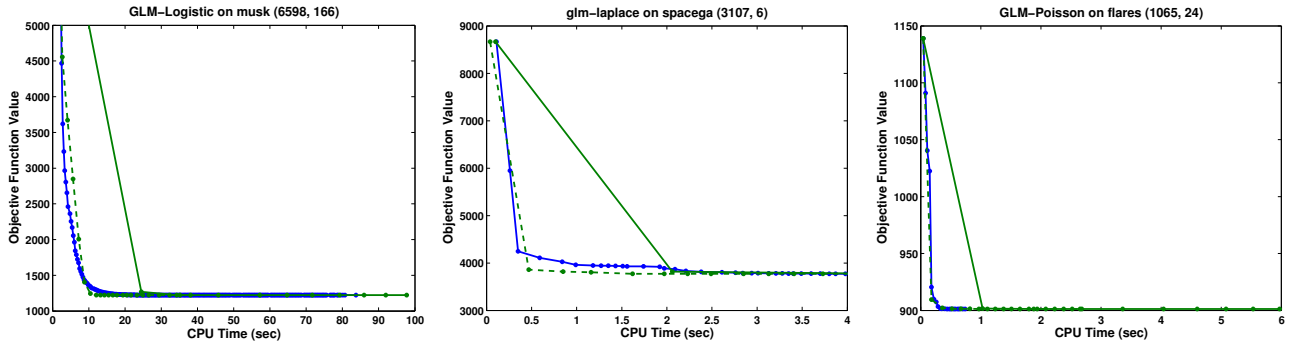


Figure 4: Evaluation for GLM showing objective function values with respect to training time. Numbers in parentheses in title refer to number of samples and dimensions of dataset. Legend for plots: GRAD (—), FPb (—). FPi (- -),

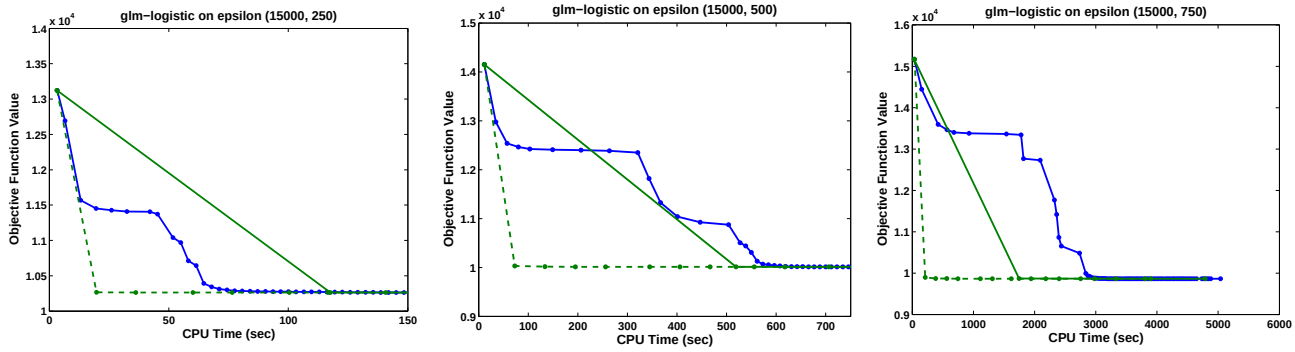


Figure 5: Evaluation for GLM showing objective function values with respect to training time for *epsilon*. The training size is fixed, but the number of features is varied. Legend for plots: GRAD (—), FPb (—). FPi (- -),