# A Fixed-Point Operator for Inference in Variational Bayesian Latent Gaussian Models

**Rishit Sheth**
rishit.sheth@tufts.edu

**Roni Khardon**
roni@cs.tufts.edu

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

## Abstract

Latent Gaussian Models (LGM) provide a rich modeling framework with general inference procedures. The variational approximation offers an effective solution for such models and has attracted a significant amount of interest. Recent work proposed a fixed-point (FP) update procedure to optimize the covariance matrix in the variational solution and demonstrated its efficacy in specific models. The paper makes three contributions. First, it shows that the same approach can be used more generally in extensions of LGM. Second, it provides an analysis identifying conditions for the convergence of the FP method. Third, it provides an extensive experimental evaluation in Gaussian processes, sparse Gaussian processes, and generalized linear models, with several non-conjugate observation likelihoods, showing wide applicability of the FP method and a significant advantage over gradient-based optimization.

## 1 INTRODUCTION

Latent Gaussian Models (LGM) provide a rich modeling framework with general inference procedures and have attracted a significant amount of interest. As argued in previous work (Challis & Barber, 2013; Khan et al., 2013), LGM capture many existing models as special cases including Gaussian Processes (GP), generalized linear models, probabilistic PCA, and more. With a small extension, LGM also capture the sparse GP model which enables efficient inference reducing the cubic complexity of standard GP.

Appearing in Proceedings of the $19^{th}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 51. 

The extended LGM model is specified by (1) below where $w \in R^d$, $f \in R^n$, and the site potentials $\phi_i(f_i)$ implicitly capture the non-Gaussian data likelihood. The distribution $p(f|w)$ varies according to the model.

$$w \sim \mathcal{N}(\mu, \Sigma), \quad f|w \sim p(f|w), \quad p(\text{data}|f) = \prod_i \phi_i(f_i), \tag{1}$$

For example, in logistic regression $w$ is the weight vector, $p(f|w) = \delta(f - H^T w)$ (where $H = (h_1, h_2, \ldots, h_n)$ and $h_i$ is the $i$th example), and $\phi_i(f_i) = \sigma(y_i f_i)$ where the label $y_i$ is in $\{-1, 1\}$. Other linear models can be specified by replacing the site potentials $\phi_i(f_i)$ yielding the generalized linear model (GLM). For Gaussian processes with mean function $m(\cdot)$ and covariance $K(\cdot,\cdot)$, we have $d = n$, and $w$ is the latent function at sample points $x$ yielding $w \sim \mathcal{N}(m(x), K(x, x^T))$. Here we have $f = w$, which for uniformity we write as $p(f|w) = \delta(f - H^T w)$ with $H = I$, and $\phi_i(f_i) = p(y_i|f_i)$ is the likelihood of observations. In the sparse model, $w$ represents the latent function at the pseudo inputs $u$ and $f$ is the latent function at $x$ (Titsias, 2009). In this case $p(f|w)$ is a linear conditional Gaussian and $\phi_i(f_i) = p(y_i|f_i)$ (Sheth et al., 2015).

Since all these models include non-conjugate priors, computation of the posterior and marginal likelihood are challenging and various approaches and approximations have been developed. Among these, the variational approach has been extensively investigated recently, as it provides a well justified criterion – maximizing a lower bound on the marginal likelihood, known as the variational lower bound (VLB). In addition, the variational approximation is computationally stable and yields good results in practice (Opper & Archambeau, 2009; Titsias, 2009; Lázaro-gredilla & Titsias, 2011; Khan et al., 2012; Challis & Barber, 2013; Khan et al., 2013; Hensman et al., 2013; Khan, 2014; Titsias & Lázaro-gredilla, 2014; Gal et al., 2014; Hensman et al., 2015; Sheth et al., 2015; Hoang et al., 2015). In this approach, a variational distribution

$$q(w, f) = q(w)p(f|w), \quad \text{where } q(w) \sim \mathcal{N}(\boldsymbol{m}, V) \tag{2}$$

is used to approximate the posterior and $(\boldsymbol{m}, V)$ are chosen to minimize the KL divergence to the true posterior. Thus the variational distribution is not in general form. It assumes a Gaussian distribution over $w$ and uses an explicit form equating $q(f|w) = p(f|w)$.

This optimization, especially optimizing $V$, is non trivial and several approaches have been proposed. In the context of GP with non-conjugate likelihoods, Opper & Archambeau (2009) observed that $V$ has a special structure and proposed a re-parametrization that reduces the number of parameters from $O(n^2)$ to $O(n)$. The work of Khan et al. (2012) showed that this parameterization is not concave and proposed a concave improvement for that algorithm. For LGM, Challis & Barber (2013) showed that for log-concave site potentials the variational lower bound is concave in $\boldsymbol{m}$ and the Cholesky factor of $V$ and proposed gradient based optimization. The work of (Khan et al., 2013; Khan, 2014) uses dual decomposition to obtain faster inference. The sparse GP model was recently explored by several groups. Here the objective was optimized with gradient search and the dual method of Khan et al. (2013) and was further developed for big data with stochastic gradients and parallelization (Hensman et al., 2013; Titsias & Lázaro-gredilla, 2014; Hensman et al., 2015; Gal et al., 2014; Sheth et al., 2015).

This paper is motivated by recent proposals to use a fixed point (FP) update, of the form $V \leftarrow T(V)$, for inference of the covariance function $V$ in some special cases of LGM. In particular, Honkela et al. (2010) proposed FP as a heuristic to simplify the update for a Gaussian covariance within their adaptation of the conjugate gradient algorithm to use natural gradients, and Sheth et al. (2015) proposed a similar update in the context of sparse GP. The work of Sheth et al. (2015) provided some empirical evidence that $T(V)$ acts as a contraction in many cases and that it often leads to fast convergence of the sparse model. But neither work provides an analysis of whether and under what conditions such an update is guaranteed to converge. Similarly we are not aware of any systematic investigation of the convergence of FP in practice across different models and site potentials.

This paper makes three contributions. The first is in observing that the FP algorithm is more widely applicable and that it can be used in the extended LGM model. The second contribution is an analysis providing sufficient conditions for convergence of the update operator $T(V)$ and showing that the convergence conditions hold for many instances of that model. The conditions for convergence rely only on properties of the site potential functions and can be tested in advance for any concrete model. The third contribution is an experimental evaluation in GP, sparse GP, and GLM for several likelihood functions showing that the FP method is widely applicable and that it offers significant advantages in convergence time over gradient based methods across all these models.

## 2 FIXED POINT UPDATES

In this section we review the variational approach to the extended LGM and the resulting FP update. Most of the development in this section is either directly stated or is implicit in previous work (Challis & Barber, 2013; Sheth et al., 2015). But the characterization of the marginal variances in (4) and corresponding implication of applicability in LGM did not previously appear in this form.

Starting with the model in (1) we can apply the variational distribution (2) to yield the following standard VLB:

$$\begin{aligned}
&\log p(\text{data}) \\
&= \log \int \mathcal{N}(w|\mu,\Sigma) p(f|w) \prod_i \phi_i(f_i) df dw \\
&\geq \int q(w,f) \log \left( \frac{\mathcal{N}(w|\mu,\Sigma) p(f|w)}{q(w,f)} \prod_i \phi_i(f_i) \right) df dw \\
&= \sum_i E_{q(w,f)}[\log \phi_i(f_i)] - d_{KL}(q(w) \| \mathcal{N}(w|\mu,\Sigma)) \\
&= \sum_i E_{q(f_i)}[\log \phi_i(f_i)] - d_{KL}(q(w) \| \mathcal{N}(w|\mu,\Sigma)) \quad (3)
\end{aligned}$$

where $d_{KL}$ is the Kullback-Leibler divergence. Below, we refer to the bound given in (3) as the VLB. Note that $p(f|w)$ does not affect the $d_{KL}$ term and it affects the first term only through the marginal distribution $q(f_i)$. In this paper we focus on cases where $q(f_i)$ is Gaussian, which holds for the models mentioned in the introduction. However, FP for the extended LGM can be used with other forms of $p(f|w)$ as long as expectations and derivatives w.r.t. $q(f_i)$, as identified below, are available or can be estimated.

Optimizing the VLB w.r.t. $\boldsymbol{m}$ is stable and can be done effectively with Newton's method or BFGS and the derivatives are given in previous work. In the following we focus on the optimization w.r.t. $V$.

To proceed, we need explicit expressions for $q(f_i)$. In the cases where $f = H^T w$, we have $f_i = h_i^T w$ and as a result $m_{q_i} = h_i^T m$, and $v_{q_i} = h_i^T V h_i$. For sparse GP, we recall that $w$ represents the latent function at the pseudo inputs $u$ and $f$ is the latent function at $x$. We therefore have that $f|w \sim \mathcal{N}(\boldsymbol{m}_x + K_{xu}K_{uu}^{-1}(w - \boldsymbol{m}_u), K_{xx} - K_{xu}K_{uu}^{-1}K_{ux})$ where we follow standard notation representing the arguments of $m(\cdot)$ and $K(\cdot,\cdot)$ using subscripts. Now, using $q(w) = \mathcal{N}(\boldsymbol{m}, V)$ and marginalizing we obtain

$m_{q_i} = m_{x_i} + K_{iu}K_{uu}^{-1}(\boldsymbol{m} - \boldsymbol{m}_u)$ and $v_{q_i} = K_{ii} + K_{iu}K_{uu}^{-1}(V - K_{uu})K_{uu}^{-1}K_{ui}$. The important point for our analysis is that in all these cases $v_{q_i}$ is a sum of a scalar and a quadratic form in $V$ and can be captured abstractly using

$$v_{q_i}(V) = c_i + d_i^T V d_i \tag{4}$$

where we emphasize the dependence on $V$. For the derivative of the VLB note that $\frac{\partial E_{q(f_i)}[\log \phi_i(f_i)]}{\partial V} = \frac{\partial E_{q(f_i)}[\log \phi_i(f_i)]}{\partial v_{q_i}} \frac{\partial v_{q_i}}{\partial V}$ and that $\frac{\partial v_{q_i}}{\partial V} = d_i d_i^T = D_i$ is a rank 1 positive semi-definite (PSD) matrix. Combining this with derivatives for $d_{KL}$, we can get two alternative expressions for the derivatives, showing that

$$\frac{\partial \text{VLB}}{\partial V} = \frac{1}{2}V^{-1} - \frac{1}{2}\Sigma^{-1} - \frac{1}{2}\sum_i \gamma_i(v_{q_i})D_i \tag{5}$$

where $\gamma_i(m_{q_i}, v_{q_i}) = -2\frac{\partial E_{q(f_i)}[\log \phi_i(f_i)]}{\partial v_{q_i}} =$

$$E_{\mathcal{N}(f_i|m_{q_i},v_{q_i})}[-\frac{\partial^2}{\partial f_i^2}\log \phi_i(f_i)] = \tag{6}$$

$$E_{\mathcal{N}(f_i|m_{q_i},v_{q_i})}[-(\frac{(f-m_{q_i})^2}{v_{q_i}} - 1)\frac{1}{v_{q_i}}\log \phi_i(f_i)] \tag{7}$$

and where (6) derived by Sheth et al. (2015) shows that $\gamma_i > 0$ for log concave site potentials and (7) derived by Challis & Barber (2013) can be used in cases where $\log \phi_i(f_i)$ is not differentiable but the expectation is differentiable.

The derivative (5) immediately suggests the FP update

$$T(V) = (\Sigma^{-1} + \sum_i \gamma_i(v_{q_i})D_i)^{-1} \tag{8}$$

The expression for $\gamma_i$ is a function of the marginal variational distribution $q(m_{q_i}, v_{q_i})$ and the generic site potentials used, and can be calculated and viewed as a function of the parameters $m_{q_i}, v_{q_i}$. Below we refer to this function abstractly as $\gamma(m, v)$ and study the convergence of the proposed method based on properties of this function.

## 3 ANALYSIS

We start by noting basic properties of the FP update. For any fixed $\boldsymbol{m}$, let $V^*$ be the optimizer of the VLB w.r.t $\boldsymbol{m}$.

**Proposition 1** *(1) $V^* = T(V^*)$, (2) $\hat{V} = T(\hat{V})$ implies that $\frac{\partial VLB}{\partial V}_{|\hat{V}} = 0$, (3) $\frac{\partial VLB}{\partial V}_{|\hat{V}} = 0$ and $\hat{V}$ is full rank implies $\hat{V} = V^*$.*

**Proof:** From (5), (8) we obviously have:

$$\frac{\partial \text{VLB}}{\partial V} = \frac{1}{2}(V^{-1} - T(V)^{-1}) \tag{9}$$

This shows that the optimal covariance $V^*$, where the derivative is zero, is a fixed-point of $T(V)$. In addition, the equation implies that if the FP method converges and $T(V) = V$ then $\frac{\partial \text{VLB}}{\partial V} = 0$ and we have reached a stationary point. Finally, it can be shown that $\frac{\partial \text{VLB}}{\partial L}L^{-1} = 2\frac{\partial \text{VLB}}{\partial V}$ where $L$ is the Cholesky factor of $V = LL^T$. Now, since for log-concave site potentials the VLB is concave in $L$ (Challis & Barber, 2013) we see that if the FP method converges to $V$ and $V$ is full rank then $V = V^*$. ■

The proposition shows that the fixed point of $T()$ identifies $V^*$. We note that a unique optimum does not imply that the VLB is concave in $V$. Next, we define sufficient conditions that guarantee that repeated application of $T()$ does converge:

**Condition 1:** for all $m, v$, $\gamma(m, v) \geq 0$.
**Condition 2a:** for all fixed values of $m$, $\gamma(m, v)$ is monotonically non-decreasing in $v$.
**Condition 2b:** for all fixed values of $m$, $\gamma(m, v)$ is monotonically non-inccreasing in $v$.

**Theorem 1** *(1) If conditions 1 and 2a hold then the FP update converges to $V^*$ or to a limit cycle of size two. (2) If conditions 1 and 2b hold then the FP update converges to $V^*$.*

**Proof:** For matrices $A, B$ we denote $A \succeq 0$ when $A$ is PSD and say that $B \succeq A$ if $B - A \succeq 0$. Now, using Condition 1 and (8) we see that for any $V$, we have $T(V)^{-1} \succeq \Sigma^{-1}$ implying that

$$\forall V, \quad T(V) \preceq \Sigma \tag{10}$$

and in particular $T(V^*) \preceq \Sigma$.

Next observe from (4) that for any $A \succeq B$, we have $v_{q_i}(A) \geq v_{q_i}(B)$, and from Condition 2a this implies $\gamma_i(v_{q_i}(A)) \geq \gamma_i(v_{q_i}(B))$. Therefore, from (8), we have

$$\forall A, B \text{ s.t. } A \succeq B, \quad T(A) \preceq T(B) \tag{11}$$

Applying (11) to $V^* \preceq \Sigma$, and using (10) to add $\Sigma$ as an upper bound we get

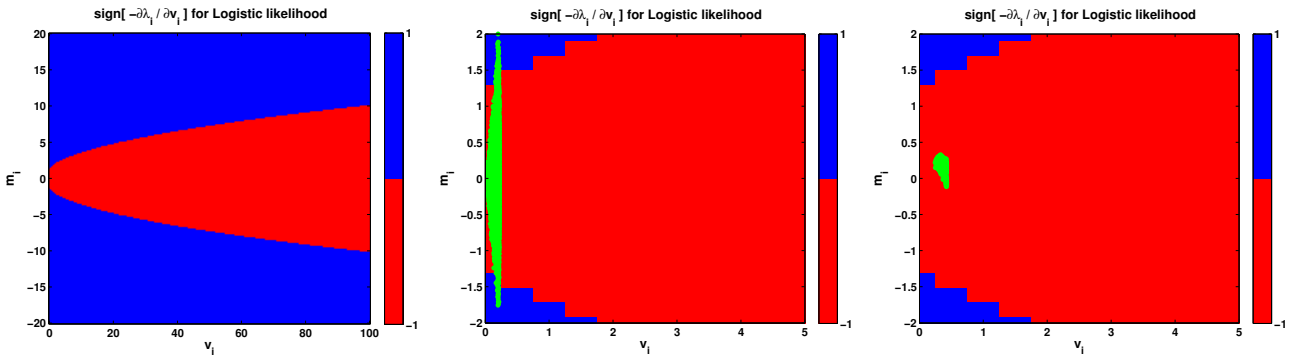
$$T(\Sigma) \preceq V^* \preceq \Sigma \tag{12}$$

Now, repeatedly applying (11) to the sequence and using (10) to add $\Sigma$ as an upper bound gives

$$T(\Sigma) \preceq T^3(\Sigma) \preceq \ldots \preceq V^* \preceq \ldots \preceq T^4(\Sigma) \preceq T^2(\Sigma) \preceq \Sigma \tag{13}$$

Denote $\gamma_i^\ell = \gamma(v_{q_i}(T^\ell(\Sigma)))$ and $\gamma_i^* = \gamma_i(v_{q_i}(V^*))$. Then Condition 2a and (13) imply

$$\gamma_i^1 \leq \gamma_i^3 \leq \ldots \leq \gamma_i^* \leq \ldots \leq \gamma_i^4 \leq \gamma_i^2 \leq \gamma_i^0 \tag{14}$$




(a) (b) (c)

Figure 1: Plots of sign($\frac{\partial}{\partial v}\gamma(m,v)$) for the logistic likelihood, with scatter plots of $(m_{q_i}, v_{q_i})$ pairs in one iteration.

As a result we see that the sequences $\{\gamma_i^{2\ell}\}$ and $\{\gamma_i^{2\ell+1}\}$ must converge because they are monotonic and bounded sequences of real numbers. This in turn implies that the sequences $\{T^{2\ell}(\Sigma)\}$ and $\{T^{2\ell+1}(\Sigma)\}$, which are defined via (8) and (14), converge. Now, if $\{\gamma_i^{2\ell}\}$ and $\{\gamma_i^{2\ell+1}\}$ converge to the same point then it must be $\gamma_i^*$ and therefore $\{T^{\ell}(\Sigma)\}$ converges to $V^*$. Otherwise, there is a gap in $\gamma$ values and $\{T^{\ell}(\Sigma)\}$ converges to a limit cycle of size two, alternating between the lower and upper bound. In summary, we have shown that under conditions 1 and 2a the FP update converges to $V^*$ or to a limit cycle of size two, i.e., $A \preceq V^* \preceq B$ and $A = T(B)$, $B = T(A)$.

We next turn to Condition 2b. In this case (10) holds but (11) is reversed to

$$\forall A, B \;\text{ s.t. }\; A \succeq B, \quad T(A) \succeq T(B)$$

yielding upon repeated application

$$\Sigma \succeq T(\Sigma) \succeq T^2(\Sigma) \succeq T^3(\Sigma) \ldots$$

In this case, $\gamma_i^{\ell}$ must converge implying that the sequence $T^{\ell}(\Sigma)$ converges to $V = T(V) = V^*$. ■

### 3.1 Applicability of the Convergence Criteria

The theorem gives sufficient conditions for convergence. We next explore when these conditions hold and cases where, although the conditions do not hold globally, weaker conditions might be sufficient in practice. We have already pointed out that Condition 1 holds for all log concave likelihoods which cover many important cases. Condition 2 holds less widely but shows an interesting structure. For some likelihoods, we have a closed form of $\gamma(m,v)$ and its derivative w.r.t. $v$ and can therefore test the condition. In particular we have:

**Remark 1** *Condition 2a holds for (1) the Poisson likelihood (with log link function)* $p(y|f) = e^{-e^f} e^{yf}/y!$ *where* $\gamma = e^{m+v/2}$*, (2) for the likelihood used in the stochastic volatility model (Rue et al., 2009; Khan et al., 2012)* $p(y|f) = \mathcal{N}(y|0, e^f)$*, where* $\gamma = e^{-2m}e^{2v}$*, and (3) for the exponential likelihood (with log link function)* $p(y|f) = e^f e^{-ye^f}$ *where* $\gamma = ye^{m+v/2}$.

When closed forms are not available we can evaluate monotonicity of $\gamma(m,v)$ for fixed $m$ empirically. Figure 1a plots sign($\frac{\partial}{\partial v}\gamma(m,v)$) for the logistic likelihood, where the color indicates regions of monotonicity, suggesting smooth behavior over large regions of the $(m,v)$ space.[1] The supplementary material includes monotonicity plots for several other likelihood functions showing similar patterns. Parts (b) and (c) show a zoomed in version of the same monotonicity plot overlaid with a scatter plot of the $(m_{q_i}, v_{q_i})$ pairs at the beginning of the second FP update, for a GLM (b) and GP model (c) on one dataset, taken from the experiments in the next section. We see that Condition 2 holds for this instance of GP (Condition 2b holds in this iteration) but not for the instance of GLM where some of the $\gamma$ values are increasing and some are decreasing. However, our experiments demonstrate that convergence does hold robustly in practice across many datasets and experimental conditions, even when such violations occur, and even in cases where Condition 1 is violated (for the Student's t likelihood).

To explore the conditions further we refer back to equation (8) which defines the FP update. Tracing the proof we see that the requirements for convergence are that the sum over the rank one matrices yields a PSD matrix and that the overall sum is increasing with respect to the ordering $\preceq$. The proof achieves this through global conditions over $\gamma(m,v)$. But the same argument goes through under the aggregate condition over $\sum \gamma_i D_i$. The aggregate condition can be

[1] Calculated on a (200x200) grid in $(m,v)$ space where each point is computed using finite differences ($\delta$=$10^{-6}$) and where $\gamma$ is calculated using quadrature ($N_{pts}$=1000).

abstracted as a less stringent sufficient condition for convergence, but it is difficult to formalize it in a compact and crisp manner, and we have therefore opted for the global conditions above.

One can easily adjust the FP algorithm to detect violations of these conditions and use a modified update. For condition 1, our implementation replaces $\gamma_i$ with 0 whenever it is negative thus maintaining the PSD condition. For violations of condition 2, one could resort to standard gradient update whenever this occurs, for example in the case illustrated for GLM. However, our experimental comparison suggests that FP is significantly faster than standard gradients and therefore this will likely yield an inferior performance.

### 3.2 Discussion: FP and Gradient Search

We next show that the FP method is closely related to gradient search with a fixed step size, giving two such interpretations. For the first, recall the relation between the gradient and $T()$ expressed in (9) and rewrite $T()$ and (9) in terms of the precision matrix $Q = V^{-1}$ so that $Q$ is updated to $T(Q)$. We have

$$\frac{\partial \text{VLB}}{\partial Q^{-1}} = \frac{1}{2}(Q - T(Q)) \tag{15}$$

$$T(Q) = Q - 2\frac{\partial \text{VLB}}{\partial Q^{-1}} \tag{16}$$

As (16) shows, $T(Q)$ takes a descent step w.r.t. the gradient of the inverse instead of an ascent step w.r.t. standard gradient. Thus the FP method can be seen as an unusual gradient method with a specific step size.

The second observation (first pointed out by an anonymous reviewer) is that the FP update is a natural gradient update with step size 1. Natural gradients (Amari & Nagaoka, 2000) adapt to the geometry of the optimized function and have been demonstrated to converge faster than standard gradients in some cases. The natural gradient is a result of pre-multiplying the standard gradient by the inverse of the Fisher information matrix $I$. Recall that exponential family distributions can be alternatively described using their natural parameter $\theta$ or mean parameter $\eta$. As shown by (Sato, 2001; Hensman et al., 2012; Hoffman et al., 2013) the natural gradient with respect to $\theta$ can be derived using standard gradients with respect to $\eta$. In particular, using $\partial_N$ to denote the natural gradient, we have $\frac{\partial_N f}{\partial \theta} = I^{-1}\frac{\partial f}{\partial \theta} = \frac{\partial f}{\partial \eta}$. The corresponding natural gradient update with step size 1 is $\theta_{new} \leftarrow \theta_{old} + \frac{\partial f}{\partial \eta}$.

In our case, $\theta = (\mathbf{r}, S) = (V^{-1}\boldsymbol{m}, \frac{1}{2}V^{-1})$ and $\eta = (h, H) = (\boldsymbol{m}, -(V + \boldsymbol{m}\boldsymbol{m}^T))$ and the update for $S$ yields $\frac{1}{2}V^{-1} = \frac{1}{2}(\Sigma^{-1} + \sum_i \gamma_i(v_{q_i})D_i)$ which is identical to the FP update. The supplementary material reviews these facts, and shows in addition that the analysis applies whenever $q(w)$ and $p(w)$ are in the same exponential family and a "FP-like" update can be derived as a natural gradient with step size 1. This also holds for the natural gradient of $\mathbf{r}$ which yields a corresponding FP update for $\boldsymbol{m}$.

It is known (Hoffman et al., 2013; Sato, 2001) that in some cases (exponential family likelihoods with conjugate priors and conjugate complete conditionals) size 1 natural gradients are equivalent to coordinate ascent optimization and they therefore converge. However, to our knowledge no existing prior analysis implies the convergence of the FP update as proved above. Specifically, the conditions required by (Hoffman et al., 2013; Sato, 2001) do not hold for the extended LGM. Moreover, exploratory experiments (provided in the supplementary material) show that applying the same type of "FP-like" update to $\boldsymbol{m}$ is not generally stable and can converge to an inferior local maximum, illustrating that no such general conditions hold. Therefore, our analysis can be seen to identify specific conditions under which size 1 natural gradients lead to convergence. It would be interesting to explore more general conditions under which convergence holds.

## 4 EXPERIMENTS

To show wide applicability, we evaluate the FP method across several probabilistic models and likelihood functions. In particular, we evaluate FP on GLM, GP, and sparse GP. We compare the performance of FP to the gradient based optimization of Challis & Barber (2013) which we denote below by GRAD. Since we are mainly concerned with the optimization and its speed, the criterion in our comparison is the value of the VLB obtained by the methods as a function of time.

Our experiments include the Poisson likelihood which satisfies Conditions 1 and 2a, the Laplace and logistic likelihoods which satisfy Condition 1 but not 2, and the Student's t likelihood which violates both conditions. In the latter case, we modify the implementation so that whenever Condition 1 is violated, i.e., $\gamma_i < 0$, it is set to zero. This heuristic ensures the positive-definiteness of the variational covariance for all fixed point iterations. Thus we test if convergence holds even when the conditions are not satisfied.

For all experiments we used the `vgai` package (Challis & Barber, 2011) that implements the GRAD method (Challis & Barber, 2013). GRAD optimizes the mean and Cholesky decomposition of the covariance jointly with L-BFGS. To facilitate as close a comparison as possible, the implementation of the fixed point methods uses `vgai` as well but replaces the optimization function call with the corresponding updates. The op-

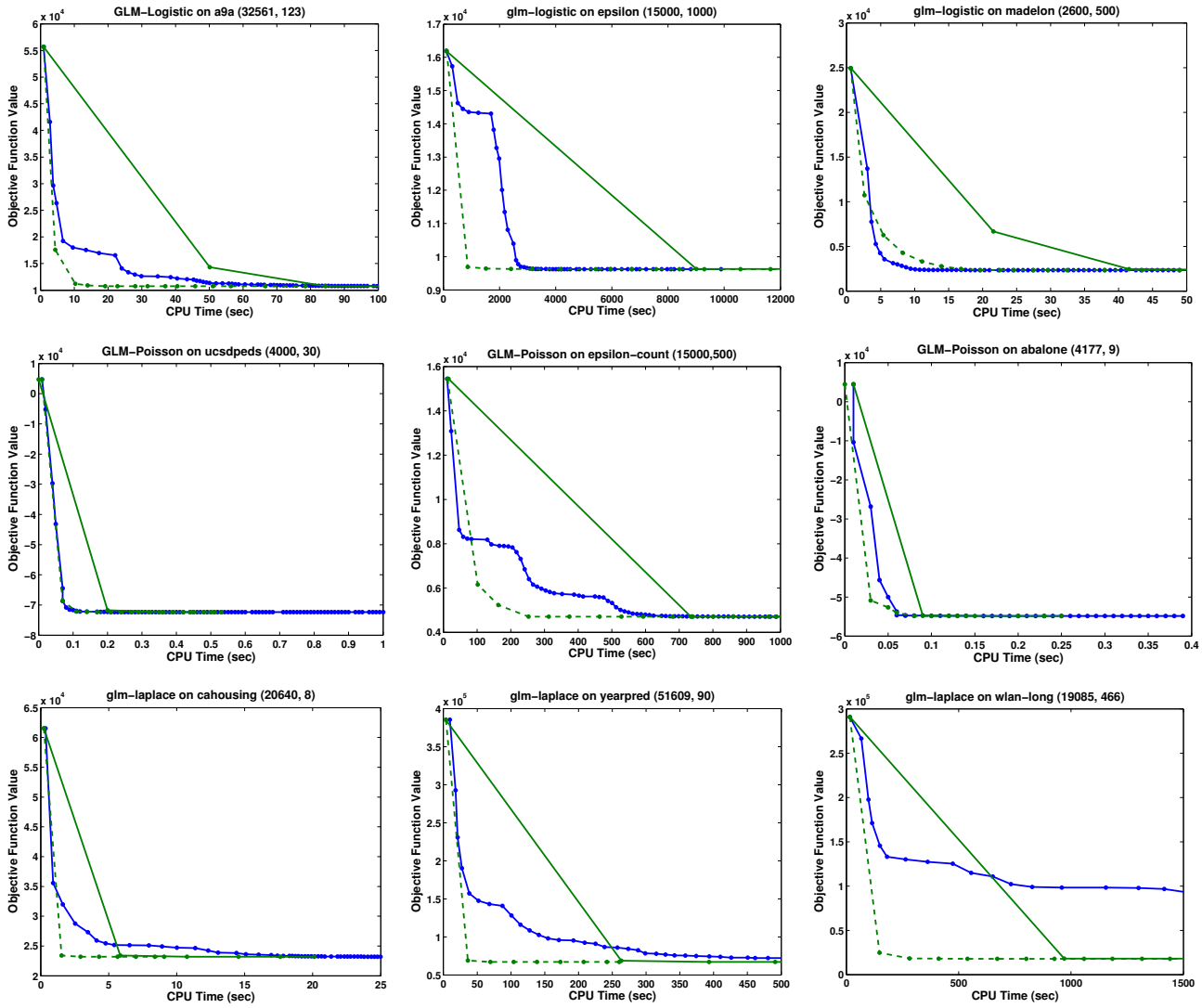



Figure 2: Evaluation for GLM showing objective function values with respect to training time. Numbers in parentheses in title refer to number of samples and dimensions of dataset. Legend for plots: GRAD (—), FPb (—). FPi (- -),

timization parameters for all algorithms were set to the same values where applicable (see supplementary material). Our experiments include two variants of the FP method. The first, as in the analysis, alternates between optimizing $\boldsymbol{m}$ and $V$ where $\boldsymbol{m}$ is optimized using Newton's method and $V$ is optimized with FP updates. This was the method recommended by Sheth et al. (2015). We have found, however, that complete optimization of $\boldsymbol{m}$ during the early iterations can be expensive, and have therefore implemented a second variant, closer to the simultaneous optimization of $(\boldsymbol{m}, V)$ performed in GRAD. In particular, the algorithm alternates between taking one gradient step for $\boldsymbol{m}$ using Newton's method, and one fixed-point update $T(V)$. When $\boldsymbol{m}$ has converged we get fixed point iterations on $V$ and similarly if $V$ has converged we get second-order optimization on $\boldsymbol{m}$. To distinguish the methods we refer to them below as FPbatch (FPb) and FPincremental (FPi).

In our experiments we have observed the cycling behavior suggested by the analysis in intermediate iterations of FPb (see supplementary material). Note, however, that even if this occurs, once $\boldsymbol{m}$ is updated in the next iteration the fixed point update for $V$ is able to exit this condition. Empirically, in all our experiments FPb and FPi do approach the optimal VLB.

The supplementary material includes a list of all datasets used in the experiments. Briefly, we selected medium size datasets to start with and added large ones to demonstrate performance in GLM. We used Z-score normalization for all features in all datasets.

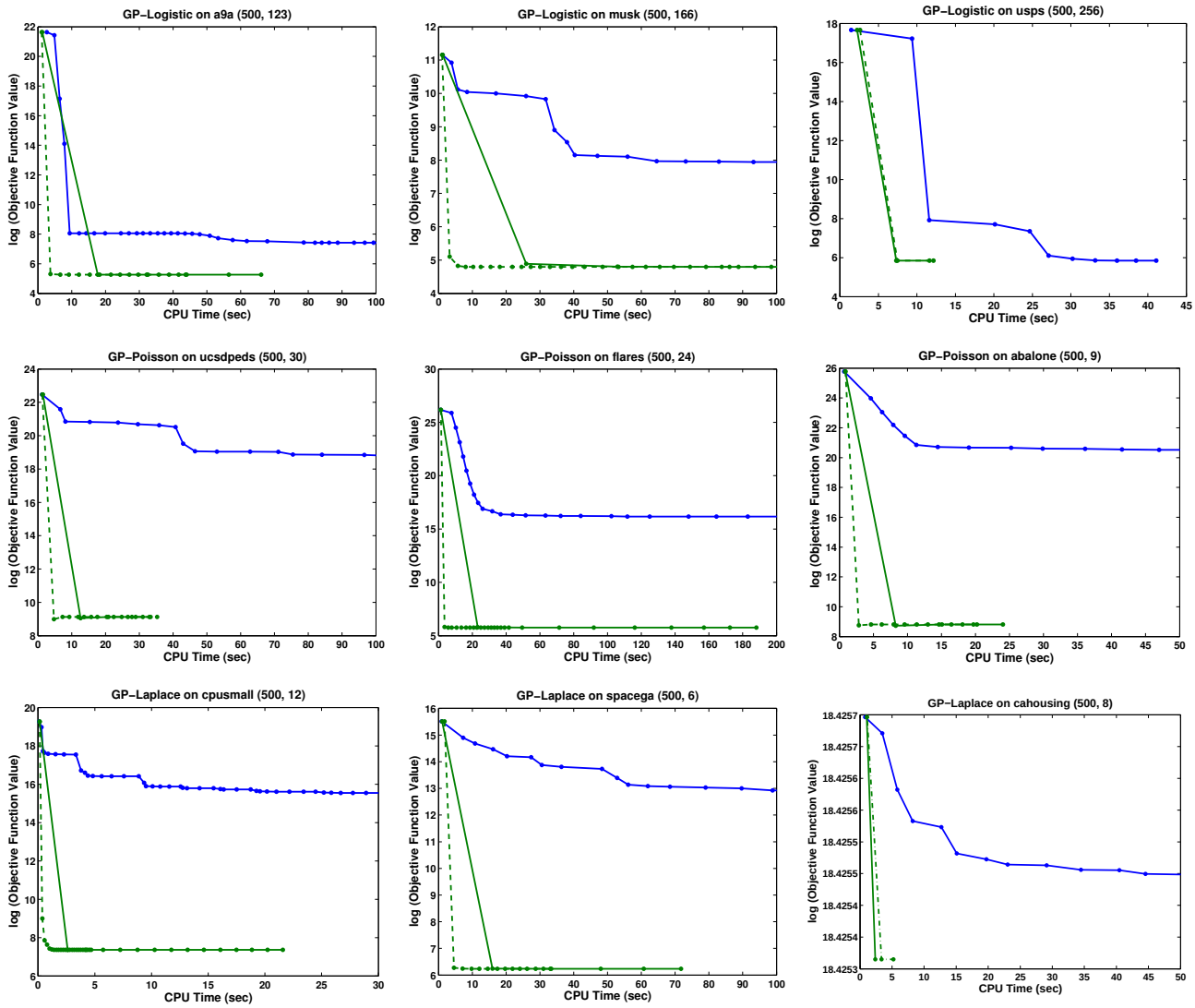



Figure 3: Evaluation for GP. See Figure 2 for description and legend.

The experimental setting is as follows. For GLM, $\Sigma = 0.1I$ in all cases except *flares* where $\Sigma = 0.001I$. For GP and sparse GP, we use a zero mean prior and $\Sigma$ is the corresponding kernel matrix, where we use the Gaussian RBF kernel with length scale and scale factor estimated from 200 randomly chosen samples (using the GPML toolbox (Rasmussen & Nickisch, 2013)). The variance of the Laplace likelihood was also estimated in this way. The logistic and Poisson likelihoods do not have parameters. The parameters of the Student's t likelihood were fixed to $\nu = 3$ and $\sigma^2 = \frac{1}{3}$. Since our focus is on the optimization procedure, hyperparameters remain fixed during the experiments and equal across the algorithms. Datasets for the GP experiment were sub-sampled to 500 samples. The inducing set for the sparse GP experiment was 100 samples randomly chosen and fixed across algorithms. The initial conditions for the optimization were set to $\boldsymbol{m} = 0$ and $V = I$, except in the case of count data (where the log link function is sensitive w.r.t. numerical stability) where $V = 0.1I$.

Note that we test several probabilistic models and several likelihoods under the same algorithmic setup for FP. This provides a robust evaluation of the FP method showing that it works well across all these cases without specific adjustment for each case.

Figure 2 shows results of experiments with GLM across classification, count regression and robust regression. The plots show a significant advantage of FPi in all cases, with GRAD reducing VLB well initially but slowing considerably thereafter in many cases. Preliminary experiments with smaller datasets (see supplementary material) showed the FP was competitive with GRAD on the GLM model but did not show a significant difference. This shows that the advantage of

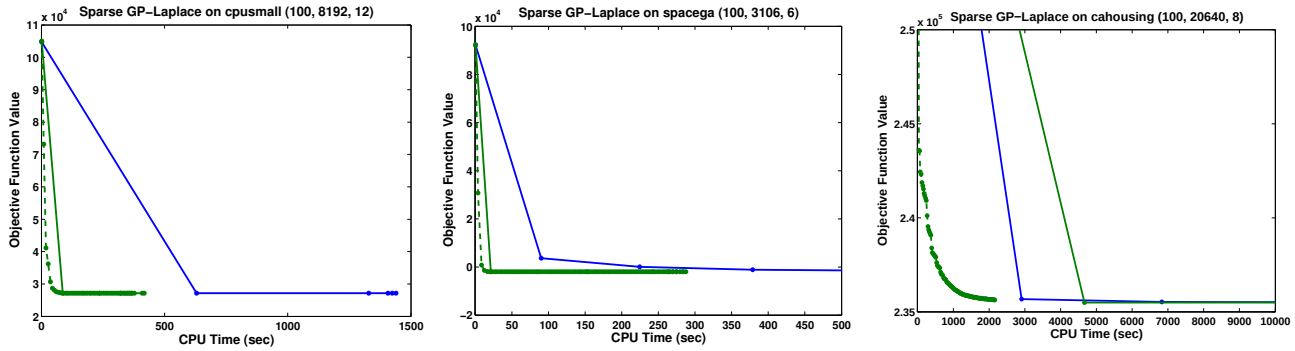



Figure 4: Evaluation for sparse GP with Laplace likelihood. See Figure 2 for legend.

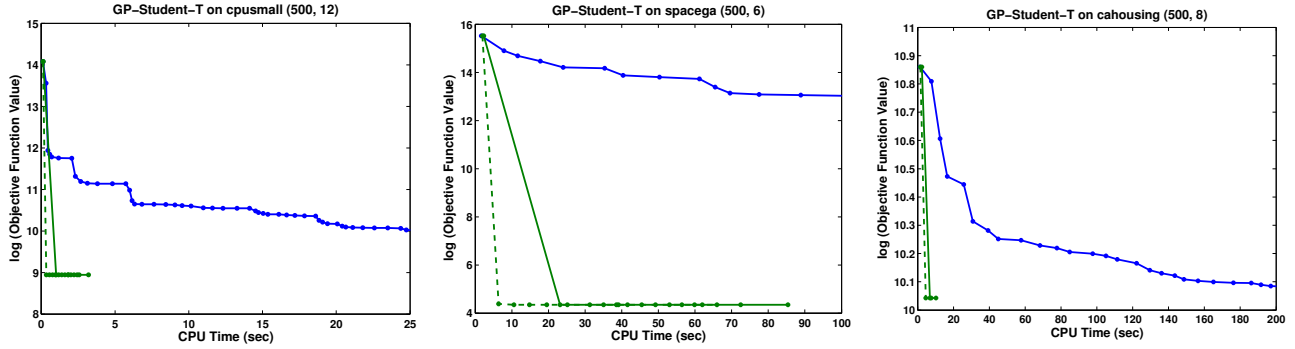



Figure 5: Evaluation for GP with (non-log-concave) Student's t likelihood. See Figure 2 for legend.

the FP method is more pronounced for larger datasets.

Figure 3 shows results of experiments with GP on the same likelihoods where both FP methods have a significant advantage over GRAD and where the differences are more dramatic than in GLM.

The FP method for the sparse GP model has already been evaluated with Poisson, logistic, and ordinal likelihoods (Sheth et al., 2015). Figure 4 complements this and shows the results of experiments with sparse GP for the Laplace likelihood. We observe the same general behavior as in the full GP model with FPi being the fastest and GRAD and FPb being slower.

Figure 5 shows results of experiments with GP with the Student's t likelihood. In this case both conditions 1 and 2 from our analysis do not hold. Nonetheless, the methods behave quite similarly in this case as well.

## 5 CONCLUSIONS

The paper shows that the FP method is applicable in the extended LGM model, provides an analysis that establishes the convergences of the FP method and provides an extensive experimental evaluation demonstrating that the FP method is applicable across GP, sparse GP, and GLM, that it converges for various likelihood functions, and that it significantly outperforms gradient based optimization in all these models.

We conclude with two directions for future work. As mentioned above, this paper focused on the case where $p(f|w)$ is linear Gaussian but the same approach is applicable as long as $q(f_i)$ and quantities relative to this distribution are efficiently computable. Probabilistic matrix factorization (Salakhutdinov & Mnih, 2008) is a LGM where $p(f|w)$ is more complex and where variational solutions have been investigated (Lim & Teh, 2007; Seeger & Bouchard, 2012). It would be interesting to develop efficient FP updates for this model. Along a different dimension, recent work on variational inference has demonstrated the utility of stochastic gradient optimization (SG), with or without natural gradients, for scalability to very large datasets. SG typically requires careful control of decreasing learning rates, whereas the FP method was shown to be related to gradient step with a fixed step size. It would be interesting to develop algorithms that combine the benefits of both methods, by incorporating FP updates within SG.

## Acknowledgments

Some of the experiments in this paper were performed on the Tufts Linux Research Cluster supported by Tufts Technology Services.

## References


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

Challis, Edward and Barber, David. `vgai` software package, 2011. `mloss.org/software/view/308/`.

Challis, Edward and Barber, David. Gaussian Kullback-Leibler Approximate Inference. *Journal of Machine Learning Research*, 14:2239–2286, 2013.

Gal, Yarin, van der Wilk, Mark, and Rasmussen, Carl. Distributed Variational Inference in Sparse Gaussian Process Regression and Latent Variable Models. In *Advances in Neural Information Processing Systems 27*, pp. 3257–3265. 2014.

Hensman, James, Rattray, Magnus, and Lawrence, Neil D. Fast Variational Inference in the Conjugate Exponential Family. In *Advances in Neural Information Processing Systems 25*, pp. 2888–2896. 2012.

Hensman, James, Fusi, Nicolo, and Lawrence, Neil D. Gaussian Processes for Big Data. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, pp. 282–290, 2013.

Hensman, James, Matthews, Alexander, and Ghahramani, Zoubin. Scalable Variational Gaussian Process Classification. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pp. 351–360, 2015.

Hoang, Trong Nghia, Hoang, Quang Minh, and Low, Bryan Kian Hsiang. A Unifying Framework of Anytime Sparse Gaussian Process Regression Models with Stochastic Variational Inference for Big Data. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 569–578, 2015.

Hoffman, Matthew D., Blei, David M., Wang, Chong, and Paisley, John William. Stochastic Variational Inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Honkela, Antti, Raiko, Tapani, Kuusela, Mikael, Tornio, Matti, and Karhunen, Juha. Approximate Riemannian Conjugate Gradient Learning for Fixed-Form Variational Bayes. *Journal of Machine Learning Research*, 11:3235–3268, 2010.

Khan, Mohammad E. Decoupled Variational Gaussian Inference. In *Advances in Neural Information Processing Systems 27*, pp. 1547–1555, 2014.

Khan, Mohammad E., Mohamed, Shakir, and Murphy, Kevin P. Fast Bayesian Inference for Non-Conjugate Gaussian Process Regression. In *Advances in Neural Information Processing Systems 25*, pp. 3149–3157. 2012.

Khan, Mohammad E., Aravkin, Aleksandr Y., Friedlander, Michael P., and Seeger, Matthias. Fast Dual Variational Inference for Non-Conjugate LGMs. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 951–959, 2013.

Lázaro-gredilla, Miguel and Titsias, Michalis K. Variational Heteroscedastic Gaussian Process Regression. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 841–848, 2011.

Lim, Y. J. and Teh, Y. W. Variational Bayesian Approach to Movie Rating Prediction. In *Proceedings of KDD Cup and Workshop*, 2007.

Opper, Manfred and Archambeau, Cédric. The Variational Gaussian Approximation Revisited. *Neural Computation*, 21(3):786–792, 2009.

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

Rue, Havard, Martino, Sara, and Chopin, Nicolas. Approximate Bayesian Inference for Latent Gaussian Models by using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

Salakhutdinov, Ruslan and Mnih, Andriy. Bayesian Probabilistic Matrix Factorization using Markov Chain Monte Carlo. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 880–887, 2008.

Sato, Masa-aki. Online Model Selection Based on the Variational Bayes. *Neural Computation*, 13(7):1649–1681, 2001.

Seeger, Matthias W. and Bouchard, Guillaume. Fast Variational Bayesian Inference for Non-Conjugate Matrix Factorization Models. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, pp. 1012–1018, 2012.

Sheth, Rishit, Wang, Yuyang, and Khardon, Roni. Sparse Variational Inference for Generalized Gaussian Process Models. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1301–1311, 2015.

Titsias, Michalis. Variational Learning of Inducing Variables in Sparse Gaussian Processes. In *the 12th International Conference on Artificial Intelligence and Statistics*, pp. 567–574, 2009.

Titsias, Michalis and Lázaro-gredilla, Miguel. Doubly Stochastic Variational Bayes for non-Conjugate Inference. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 1971–1979, 2014.

# 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 $\eta_q$ 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, $\theta$ and $\eta$ (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 $\eta_q$, 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 $\eta_q$

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

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

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 $\gamma$s 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.

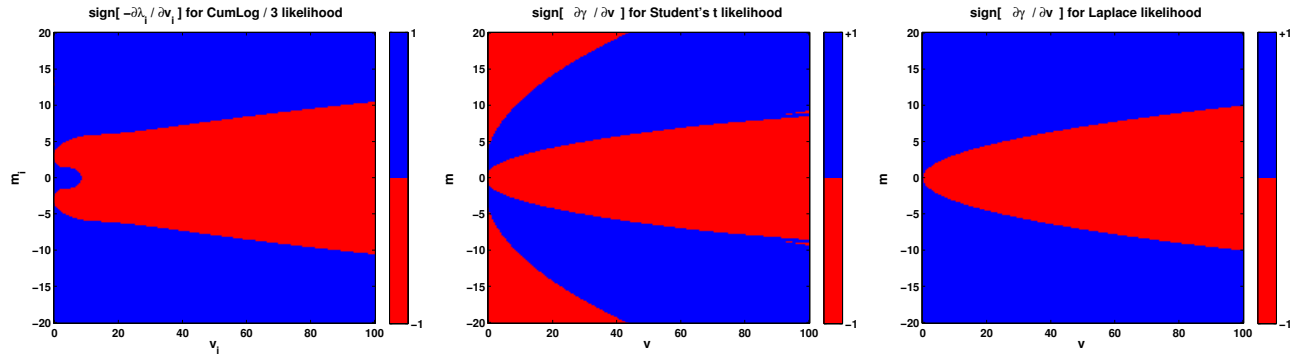
Guyon, Isabelle, Gunn, Steve, Ben-Hur, Asa, and Dror, Gideon. Result analysis of the NIPS 2003 feature selection challenge. In *Advances in neural information processing systems*, pp. 545–552, 2004.

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.

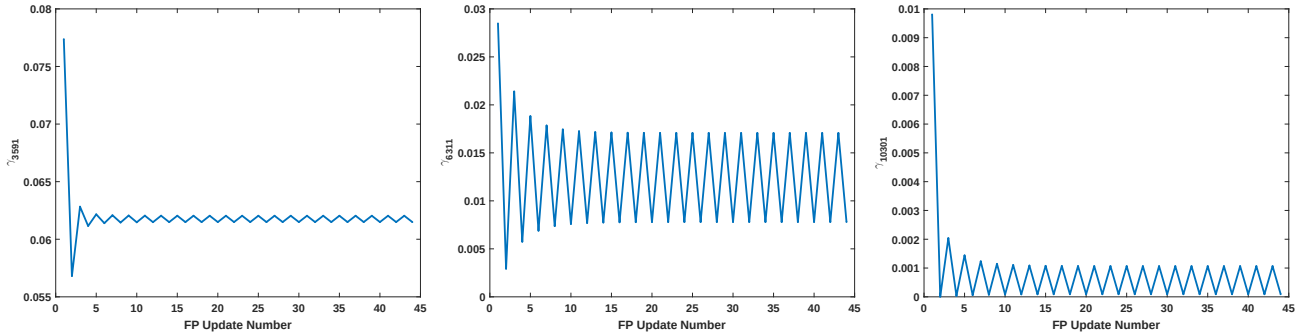



Figure 2: Plot of $\gamma_i$ for $i = 3591, 6311$, and 10301 (out of 19085) vs. FP update number.

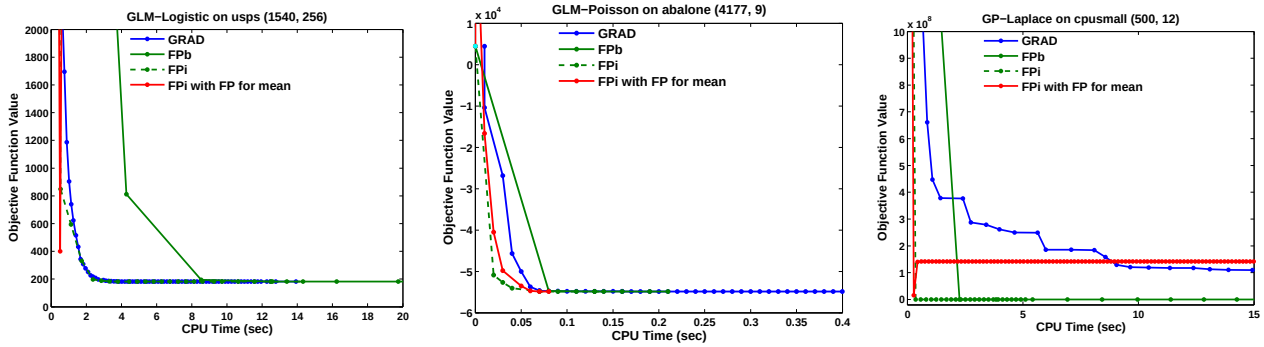



Figure 3: Comparison of using FP for the mean against other methods for three datasets.

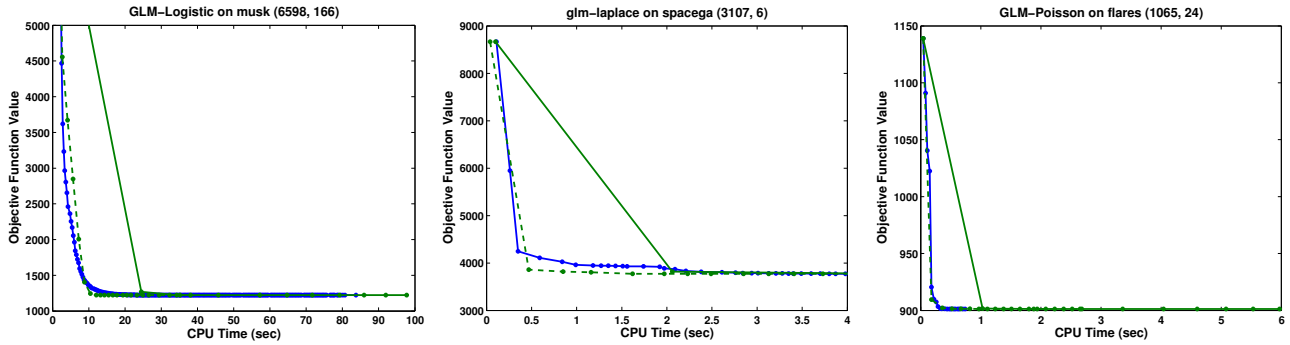



Figure 4: Evaluation for GLM showing objective function values with respect to training time. Numbers in parentheses in title refer to number of samples and dimensions of dataset. Legend for plots: GRAD (—), FPb (—). FPi (- -),

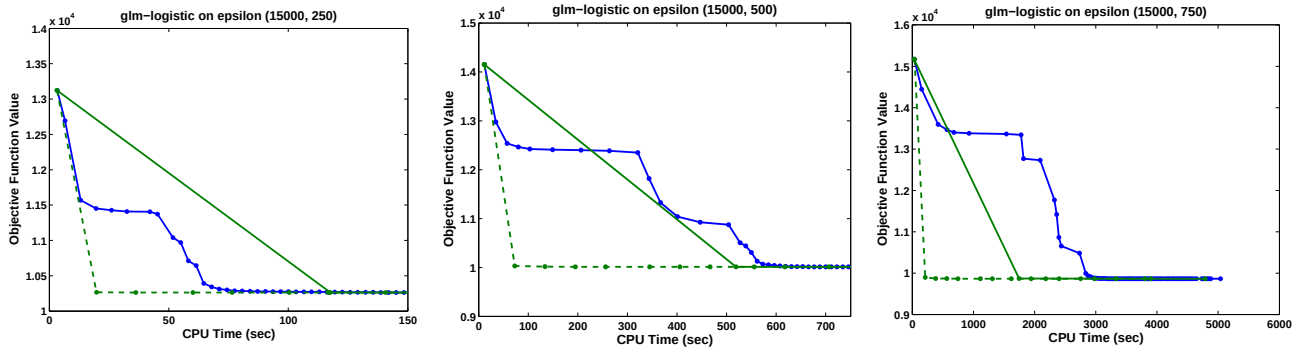



Figure 5: Evaluation for GLM showing objective function values with respect to training time for *epsilon*. The training size is fixed, but the number of features is varied. Legend for plots: GRAD (—), FPb (—). FPi (- -),