The Abdus Salam
International Centre
for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

2361-7

School on Large Scale Problems in Machine Learning and Workshop on Common Concepts in Machine Learning and Statistical Physics

20 - 31 August 2012

Large Scale Variational Bayesian Inference for Continuous Variable Models - Exercise Sheet

Matthias SEEGER
Laboratory for Probabilistic Machine Learning, EPFL, CH-1015 Lausanne
Switzerland

ICTP School on Large Scale Problems in Machine Learning: Large Scale Variational Bayesian Inference for Continuous Variable Models — Exercise Sheet

Matthias Seeger
Probabilistic Machine Learning Laboratory
Ecole Polytechnique Fédérale de Lausanne
INR 112, Station 14, CH-1015 Lausanne
matthias.seeger@epfl.ch

August 21, 2012

1 Super-Gaussian Bounding for Laplace Potentials

Recall the super-Gaussian bounding variational relaxation of Bayesian inference and the coordinate update algorithm from the course. In this exercise, you will work out the details of a single update for a sparse linear model with Laplace potentials:

$$t_j(s_j) = e^{\tau|s_j|}, \quad \tau > 0. \tag{1}$$

We omit the normalization constant $\tau/2$ of the Laplace density, as it is not important in the context here. The posterior distribution is

$$P(\boldsymbol{u}|\boldsymbol{y}) = Z^{-1} P(\boldsymbol{y}|\boldsymbol{u}) \prod_{j=1}^{q} t_j(s_j), \quad \boldsymbol{s} = \boldsymbol{B}\boldsymbol{u}, \quad P(\boldsymbol{y}|\boldsymbol{u}) = N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{u}, \sigma^2\boldsymbol{I}). \tag{2}$$

However, the detailed components of the model do not matter much in this exercise.

a) Recall that Laplace potentials are super-Gaussian:

$$t_j(s_j) = \max_{\gamma_j \geq 0} e^{-\frac{1}{2}s_j^2/\gamma_j - \frac{1}{2}h_j(\gamma_j)}, \quad h_j(\gamma_j) = \tau^2\gamma_j.$$

Review the material from the lecture slides and confirm the result for $h_j(\gamma_j)$, given the definition

$$h_j(\gamma_j) = \max_{x \geq 0} -x/\gamma_j - 2\log t_j(\sqrt{x})$$

b) The criterion to be minimized w.r.t. $\boldsymbol{\gamma} = [\gamma_j]$ is

$$\phi(\boldsymbol{\gamma}) = -2\log Z_Q + \sum_{j=1}^{q} h_j(\gamma_j), \quad Z_Q = \int P(\boldsymbol{y}|\boldsymbol{u}) \prod_{j=1}^{q} e^{-\frac{1}{2}s_j^2/\gamma_j}\, d\boldsymbol{u}.$$

Here, the Gaussian approximation is

$$Q(\boldsymbol{u}|\boldsymbol{y}) = Z_Q^{-1} P(\boldsymbol{y}|\boldsymbol{u}) \prod_{j=1}^{q} e^{-\frac{1}{2} s_j^2/\gamma_j}. \tag{3}$$

Prove that

$$\frac{\partial}{\partial \gamma_j^{-1}} - 2\log Z_Q = \mathrm{E}_Q[s_j^2],$$

where $\mathrm{E}_Q[\cdot]$ denotes expectation over $Q(s_j|\boldsymbol{y})$, a marginal of $Q(\boldsymbol{u}|\boldsymbol{y})$.

c) Derive an update equation for γ_j by setting the derivative $\partial\phi/\partial\gamma_j$ equal to zero.

d) [advanced] Even though the update equation you derived in the previous part works well and is commonly used, it does not necessarily minimize $\phi(\boldsymbol{\gamma})$ w.r.t. γ_j completely. Why? How would you do the complete minimization, without having to recompute the marginal $Q(s_j|\boldsymbol{y})$ during the iteration?

2 Gaussian KL Minimization and Super-Gaussian Bounding

In this exercise, we assume that the non-Gaussian potentials $t_j(s_j)$ are even and super-Gaussian. Recall the following two variational inference relaxations, applied to the posterior distribution (2). First, super-Gaussian bounding, based on the bound

$$-2\log Z \le \min_{\boldsymbol{\gamma}} \left\{ \phi_{\mathrm{SG}} = -2\log Z_Q + \sum_{j=1}^{q} h_j(\gamma_j) \right\}.$$

Here, $\log Z_Q$ is the partition function for the Gaussian posterior approximation (3), and

$$t_j(s_j) = \max_{\gamma_j \ge 0} e^{-\frac{1}{2}(s_j^2/\gamma_j + h_j(\gamma_j))}. \tag{4}$$

Second, Gaussian KL minimization, based on the bound

$$-2\log Z \le \min_{\boldsymbol{\gamma},\boldsymbol{b}} \left\{ \phi_{\mathrm{KL}} = -2\log Z_Q + \sum_{j=1}^{q} 2\mathrm{E}_Q\left[-\log t_j(s_j) - s_j^2/(2\gamma_j) + b_j s_j\right] \right\}.$$

Here, $\mathrm{E}_Q[\cdot]$ denotes expectation over $Q(s_j|\boldsymbol{y})$, the marginal of $Q(\boldsymbol{u}|\boldsymbol{y})$.

a) Prove that Gaussian KL minimization provides a tighter bound to $-2\log Z$ than super-Gaussian bounding (at their respective optimum points):

$$\min_{\boldsymbol{\gamma},\boldsymbol{b}} \phi_{\mathrm{KL}}(\boldsymbol{\gamma},\boldsymbol{b}) \le \min_{\boldsymbol{\gamma}} \phi_{\mathrm{SG}}(\boldsymbol{\gamma})$$

Hint: If $\boldsymbol{\gamma}_* = \operatorname{argmin} \phi_{\mathrm{SG}}(\boldsymbol{\gamma})$, show that

$$\phi_{\mathrm{KL}}(\boldsymbol{\gamma}_*, \mathbf{0}) \le \phi_{\mathrm{SG}}(\boldsymbol{\gamma}_*).$$

3 Efficient Parameterization of Gaussian KL Minimization

Recall Gaussian KL minimization from the course and from Exercise 2 above. Naturally, one would run Gaussian KL minimization over *all* Gaussians $Q(\boldsymbol{u}|\boldsymbol{y}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In this exercise, you will show that it is sufficient to optimize over Gaussians with covariance matrices which are parameterized in terms of $O(q)$ parameters, since any local optimum must have this particular form.

Consider a problem which gives rise to the posterior distribution

$$P(\boldsymbol{u}|\boldsymbol{y}) = Z^{-1} e^{-\frac{1}{2}\boldsymbol{u}^T \boldsymbol{E}\boldsymbol{u} + \boldsymbol{c}^T\boldsymbol{u}} \prod_{j=1}^{q} t_j(s_j), \quad \boldsymbol{s} = \boldsymbol{B}\boldsymbol{u}.$$

Here, $\boldsymbol{E}$ is positive semidefinite (meaning that $\boldsymbol{v}^T\boldsymbol{E}\boldsymbol{v} \geq 0$ for all $\boldsymbol{v}$), but may be singular. In our running example, $\boldsymbol{E} = \sigma^{-2}\boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{c} = \sigma^{-2}\boldsymbol{X}^T\boldsymbol{y}$. Gaussian KL minimization over *general* Gaussians is given by

$$\log Z \geq \max_{Q(\boldsymbol{u}|\boldsymbol{y}) = N(\boldsymbol{\mu},\boldsymbol{\Sigma})} \mathrm{E}_Q\left[\log \frac{e^{-\frac{1}{2}\boldsymbol{u}^T\boldsymbol{E}\boldsymbol{u}+\boldsymbol{c}^T\boldsymbol{u}} \prod_j t_j(s_j)}{Q(\boldsymbol{u}|\boldsymbol{y})}\right]. \tag{5}$$

a) Show that minus two times the lower bound in (5) can be written (up to an additive constant) as

$$\phi = -\log|\boldsymbol{\Sigma}| + \operatorname{tr}\boldsymbol{E}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T) - 2\boldsymbol{c}^T\boldsymbol{\mu} + \sum_{j=1}^{q}\nu_j. \tag{6}$$

Here, ν_j is a function of $Q(s_j|\boldsymbol{y})$, the marginal of $Q(\boldsymbol{u}|\boldsymbol{y}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
Hint: Use the definition of the differential entropy of a Gaussian $Q(\boldsymbol{u}|\boldsymbol{y}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$\mathrm{E}_Q\left[-\log Q(\boldsymbol{u}|\boldsymbol{y})\right] = \frac{1}{2}\log|2\pi e\boldsymbol{\Sigma}|.$$

b) Denote $Q(s_j|\boldsymbol{y}) = N(h_j, \rho_j)$, where $\rho_j = \boldsymbol{b}_j^T\boldsymbol{\Sigma}\boldsymbol{b}_j$, $\boldsymbol{b}_j$ the j-th row of $\boldsymbol{B}$. Compactly: $\boldsymbol{\rho} = \operatorname{diag}^{-1}(\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^T)$. We are not interested in the mean $\boldsymbol{\mu}$, consider it fixed. Suppose that $\boldsymbol{\Sigma}_*$ is a stationary point of ϕ:

$$\nabla_{\boldsymbol{\Sigma}_*}\phi = \boldsymbol{0}.$$

Show that

$$\boldsymbol{\Sigma}_* = \left(\boldsymbol{E} + \boldsymbol{B}^T(\operatorname{diag}\boldsymbol{\pi})\boldsymbol{B}\right)^{-1}, \quad \pi_j = \frac{\partial\nu_j}{\partial\rho_j}.$$

Hint: Use that $\nabla_{\boldsymbol{\Sigma}}\log|\boldsymbol{\Sigma}| = \boldsymbol{\Sigma}^{-1}$ for a symmetric nonsingular matrix $\boldsymbol{\Sigma}$ with positive determinant.

4 Coordinate Update Algorithm for Gaussian KL Minimization

Recall Gaussian KL minimization from the course and from Exercises 2 and 3 above. In particular, we will use setup and notation from the latter. Here, you will work out a coordinate update algorithm to solve this relaxation.

Throughout this exercise, we assume that $\boldsymbol{B} = \boldsymbol{I}$, so that $\boldsymbol{s} = \boldsymbol{u}$ and $q = n$. We will use the approximation family of Gaussians

$$Q(\boldsymbol{u}|\boldsymbol{y}) = N\bigl(\boldsymbol{\mu}, \underbrace{(\boldsymbol{E} + \boldsymbol{\Pi})^{-1}}_{=:\boldsymbol{\Sigma}}\bigr), \quad \boldsymbol{\Pi} = \operatorname{diag} \boldsymbol{\pi},$$

with parameters $\boldsymbol{\mu}, \boldsymbol{\pi}$ (this is slightly different from the course, where parameters were $\boldsymbol{b}, \boldsymbol{\pi}$). The algorithm iterates between updating the mean $\boldsymbol{\mu}$ by conjugate gradients, and updating $\boldsymbol{\pi}$ coordinate by coordinate. The former has been discussed in the course, we will concentrate on the latter.

a) Suppose we are to update π_j for $j \in \{1, \dots, n\}$. The marginal before the update is $Q(u_j|\boldsymbol{y}) = N(\mu_j, \rho_j)$. The goal is to minimize the criterion $\phi = \phi_{\mathrm{KL}}$ w.r.t. π_j. Denote the marginal after the update by $Q'(u_j|\boldsymbol{y}) = N(\mu'_j, \rho'_j)$, with parameter π'_j. Express π'_j in terms of ρ_j, ρ'_j.
Hint: Write $Q'(u_j|\boldsymbol{y}) \propto Q(u_j|\boldsymbol{y}) e^{-\frac{1}{2}(\Delta\pi_j)u_j^2}$, where $\Delta\pi_j = \pi'_j - \pi_j$.
Given this fact, we see that the problem of updating π_j is equivalent to minimizing ϕ w.r.t. the marginal variance ρ_j, the j-th diagonal entry of the covariance $\boldsymbol{\Sigma}$.

b) Denote the inverse covariance matrix by $\boldsymbol{P} = \boldsymbol{\Sigma}^{-1} = \boldsymbol{E} + \boldsymbol{\Pi}$. In order to update $\rho_j \to \rho'_j$, you will use an iterative method cycling between $p'_j = \boldsymbol{P}'_{jj}$ (j-th diagonal entry of $\boldsymbol{P}'$) and ρ'_j. Give an update equation for p'_j.
Hint: Use the criterion form (6) (and your results from Exercise 3), setting the derivative w.r.t. ρ'_j equal to zero.

c) Given the new value for p'_j, what is the new value for ρ'_j?

It remains to update the marginal variances $\boldsymbol{\rho}$, given that $\pi_j \to \pi'_j$. A simple idea is to maintain the covariance matrix $\boldsymbol{\Sigma}$, and to update it using the Sherman-Morrison-Woodbury formula. However, this is numerically instable. A better idea is to maintain a Cholesky factorization $\boldsymbol{P} = \boldsymbol{E} + \boldsymbol{\Pi} = \boldsymbol{L}\boldsymbol{L}^T$, and to update $\boldsymbol{L}$ given $\pi_j \to \pi'_j$. Details can be found in [2, 3], software for low rank Cholesky updates is available from my homepage. Each update costs $O(n^2)$.

5 Spectral Analysis of Conjugate Gradients Algorithm

Recall the conjugate gradients (CG) algorithm for approximately solving $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, where $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is positive definite. The algorithm minimizes the quadratic $q(\boldsymbol{x}) = (1/2)\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x}$ iteratively, constructing a sequence $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, ..., requiring a single matrix-vector multiplication with $\boldsymbol{A}$ per iteration. After at most n steps, neglecting numerical errors (which, in practice, you cannot!), the exact solution $\boldsymbol{x}_* = \boldsymbol{A}^{-1}\boldsymbol{b}$ is reached, in that $\boldsymbol{x}_n = \boldsymbol{x}_*$. Depending on $\boldsymbol{A}$, this can also happen earlier (you can use the Cayley-Hamilton theorem from linear algebra to understand this point). However, the main rationale for CG today is to *approximate* $\boldsymbol{x}_*$ by $\boldsymbol{x}_k$ with $k \ll n$. Whether $\boldsymbol{x}_k$ is close to $\boldsymbol{x}_*$ or not, depends on properties of $\boldsymbol{A}$ and $\boldsymbol{b}$ (it depends on numerical errors as well, but we ignore these in the present exercise, assuming that all computations are exact). In this exercise, we will analyze the convergence behaviour in terms of the eigenspectrum of $\boldsymbol{A}$.

Let $\boldsymbol{x}_* = \boldsymbol{A}^{-1}\boldsymbol{b}$ and $q_* = q(\boldsymbol{x}_*) = -(1/2)\boldsymbol{b}^T\boldsymbol{A}^{-1}\boldsymbol{b}$. Then, $q(\boldsymbol{x}_k)$ is nonincreasing and $\geq q_*$. We'll try to bound $q(\boldsymbol{x}_k) - q_*$. The eigendecomposition is $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$, $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ orthonormal ($\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$), $\boldsymbol{\Lambda}$ diagonal (positive elements).

a) Assume that we start from $\boldsymbol{x}_0 = \boldsymbol{0}$. The Krylov subspace $\mathcal{K}_k$ is spanned by $\{\boldsymbol{A}^j\boldsymbol{b} \,|\, j = 0, \dots, k-1\}$. We saw that $\boldsymbol{x}_k = \operatorname{argmin}_{\boldsymbol{x}\in\mathcal{K}_k} q(\boldsymbol{x})$. For a polynomial $P(t) = \sum_{j=0}^{k-1} \alpha_j t^j$, $\alpha_j \in \mathbb{R}$, define

$$P(\boldsymbol{B}) := \sum_{j=0}^{k-1} \alpha_j \boldsymbol{B}^j, \quad \boldsymbol{B} \in \mathbb{R}^{n\times n}.$$

Take care that $P(\boldsymbol{B})P(\boldsymbol{C}) \neq P(\boldsymbol{C})P(\boldsymbol{B})$ in general if $\boldsymbol{B}$, $\boldsymbol{C}$ do not commute, and recall that $\boldsymbol{B}^0 := \boldsymbol{I}$. Show that

$$P(\boldsymbol{A}) = \boldsymbol{Q}P(\boldsymbol{\Lambda})\boldsymbol{Q}^T.$$

Defining $\boldsymbol{y} = \boldsymbol{Q}^T\boldsymbol{x}$, $\bar{\boldsymbol{b}} = \boldsymbol{Q}^T\boldsymbol{b}$, show that $q(\boldsymbol{x})$ and q_* can be written in terms of $\boldsymbol{y}$, $\bar{\boldsymbol{b}}$, and $\{\lambda_i\}$.
Hint: What is the eigendecomposition of $\boldsymbol{A}^{-1}$?

b) Let $\boldsymbol{y}_k = \boldsymbol{Q}^T\boldsymbol{x}_k$. Show that $\boldsymbol{x}_k = P_k(\boldsymbol{A})\boldsymbol{b}$ for some polynomial $P_k(t)$ of degree $< k$, and that $\boldsymbol{y}_k = P_k(\boldsymbol{\Lambda})\bar{\boldsymbol{b}}$. In other words, $y_{k,i} = P_k(\lambda_i)\bar{b}_i$. By writing $q(\boldsymbol{x}_k)$ in terms of $\boldsymbol{y}_k$, prove that

$$q(\boldsymbol{x}_k) - q_* = \min_{P_k \,|\, \deg(P_k)<k} (1/2)\sum_{i=1}^{n} (\bar{b}_i^2/\lambda_i)(\lambda_i P_k(\lambda_i) - 1)^2.$$

Here, $\deg(P_k)$ is the degree of P_k, the largest j such that t^j features with a nonzero coefficient.

c) Argue that this means that the error $q(\boldsymbol{x}_k) - q_*$ is bounded in terms of a polynomial P of degree $\leq k$, such that $P(0) = -1$ (equivalent: such that $P(0) = 1$). The existence of such a polynomial which is small on *all* λ_i, implies that the error is small. Prove that if $\{\lambda_1, \dots, \lambda_n\} = \{\kappa_1, \dots, \kappa_k\}$, *i.e.* if $\boldsymbol{A}$ has no more than k different eigenvalues, then $\boldsymbol{x}_k = \boldsymbol{x}_*$.
Remark: Using Chebishev polynomials, $\min_{P_k} \max_{t\in[\lambda_{\min},\lambda_{\max}]} P_k(t)$ can be determined for $0 < \lambda_{\min} \leq \lambda_{\max}$, which leads to the worst-case error bound

$$q(\boldsymbol{x}_k) - q_* \leq \left(\frac{\rho - 1}{\rho + 1}\right)^k, \quad \rho = \sqrt{\lambda_{\max}/\lambda_{\min}}.$$

Therefore, we should aim for well-conditioned matrices $\boldsymbol{A}$ ($\lambda_{\max}/\lambda_{\min}$ small), which is what many preconditioning stategies try to do. On the other hand, if the spectrum of $\boldsymbol{A}$ comes in separate clusters, this bound is overly pessimistic.

6 Super-Gaussian Bounding for Bernoulli Potentials

In the course, we have discussed the super-Gaussian bounding inference approximation for *even* potentials only: $t(-s) = t(s)$. The Bernoulli likelihood potential violates this assumption:

$$t(s) = \frac{1}{1 + e^{-ys}}, \quad y \in \{-1, +1\}. \tag{7}$$

Super-Gaussian bounding applies to this potential, even though it is not even. Details can be found in [1, 4]. The generalized definition of a *super-Gaussian* $t(s)$ requires that

$$t(s) = \max_{\gamma \geq 0} e^{bs - \frac{1}{2}(s^2/(2\gamma) + h(\gamma))} = e^{bs} \tilde{t}(s).$$

Here, b is a constant. Different from γ, it is not optimized over. $\tilde{t}(s)$ is even and super-Gaussian.

Bernoulli potentials are used to model the likelihood $P(\boldsymbol{y}|\boldsymbol{s})$ in binary classification models, where $y_j \in \{-1, +1\}$ is the class label, $s_j = \log(P(y_j = +1|\boldsymbol{b}_j)/P(y_j = -1|\boldsymbol{b}_j))$ is the log odds ratio. For this exercise, will use a linear classification model with input points $\boldsymbol{b}_j$, class labels y_j, and classifier weights $\boldsymbol{u}$. Here, $s_j = \boldsymbol{b}_j^T \boldsymbol{u}$ (or $\boldsymbol{s} = \boldsymbol{B}\boldsymbol{u}$). There are n weights, q training datapoints. Likelihood $P(\boldsymbol{y}|\boldsymbol{u})$ and prior $P(\boldsymbol{u})$ are

$$P(\boldsymbol{y}|\boldsymbol{u}) = \prod_{j=1}^{q} t(s_j), \quad P(\boldsymbol{u}) = N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}).$$

In contrast to our sparse linear model example, the likelihood is non-Gaussian here, the prior is Gaussian.

a) Show that the Bernoulli potential (7) is super-Gaussian. What is b, what is $\tilde{t}(s)$? *Hint*: Use the fact that $x \mapsto \log\cosh(bx^{1/2})$ is a concave function for $x \geq 0$.

b) Show that both the Bernoulli potential $t(s)$ and $\tilde{t}(s)$ are log-concave (meaning that $-\log t(s)$ is convex).

c) How does the inner loop penalized least squares optimization problem look like for the binary classification model with Bernoulli likelihood? Show that this problem is convex.
Hint: It is *not* necessary to work out $h(\gamma_j)$. Instead, follow the derivation given in the course for the Laplace potential, replacing $|s_j|$ by $(z_j + s_j^2)^{1/2}$ in $\tilde{t}(s)$.

7 Proximal Map for Inner Loop Optimization Problem

Recall the double loop algorithm for super-Gaussian bounding from the course. For the sparse linear model with Laplace potentials (1), we need to solve inner loop problems of the form

$$\min_{\boldsymbol{u}_*} \sigma^{-2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}_*\|^2 + 2\tau \sum_{j=1}^{q} \sqrt{z_j + s_{*j}^2}.$$

Also, recall the alternating direction method of multipliers (ADMM), an augmented Lagrangian solver for penalized least squares problems of this form. In the course, we applied it to the MAP estimation problem, where the penalizer was $2\tau \sum_j |s_{*,j}|$, which corresponds to $z_j = 0$ above. However, for variational inference, we have that $z_j > 0$ for all j (see [4] for a proof).

In this exercise, you will work out how to solve the proximal map problem

$$s' = \mathrm{prox}(r) = \operatorname*{argmin}_s \kappa\sqrt{z + s^2} + \frac{1}{2}(s - r)^2, \quad z > 0.$$

We then obtain an algorithm to solve the inner loop problem by configuring the ADMM method discussed in the course with this primitive.

a) If $r < 0$, show that

$$\mathrm{prox}(r) = -\mathrm{prox}(-r),$$

and that $s' \geq 0$ for $r > 0$. We can restrict ourselves to $r > 0$ and optimize over $s \geq 0$ only.

b) Convince yourself that the reparameterization $s \to s/z^{1/2}$, $r \to r/z^{1/2}$, $\kappa \to \kappa/z^{1/2}$ can be used to attain $z = 1$. The problem to be solved is

$$s' = \operatorname*{argmin}_{s \geq 0} \kappa\sqrt{1 + s^2} + \frac{1}{2}(s - r)^2, \quad r > 0. \tag{8}$$

Prove that $s' \leq r$ (which means that the value r is shrunk). Also prove that

$$s' > \max\{r/(1 + \kappa), r - \kappa\}.$$

This means that $s' > 0$ for $r > 0$, in contrast to the situation for MAP estimation.
Hint: Work out the stationary equation for the optimization problem. The definition $y = (1 + s^2)^{1/2}$ may be helpful.

c) Show that the solution s' of (8) is obtained as root of a quartic equation (polynomial equation of degree 4) with real coefficients. Determine the quartic.
There is an algorithm to analytically solve for the roots of a quartic equation (see `http://en.wikipedia.org/wiki/Quartic_equation`), even though it is a bit complicated.

8 Bound on Marginal Variances

Consider the Gaussian distribution $Q(\boldsymbol{u}|\boldsymbol{y})$ from (3). We assume that all $\gamma_j > 0$.

a) Prove the bound

$$\mathrm{Var}_Q[s_j|\boldsymbol{y}] \leq \gamma_j$$

on the variance of the marginal $Q(s_j|\boldsymbol{y})$.
Hint: Use the identity

$$\boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{v} = \max_{\boldsymbol{x}} 2\boldsymbol{v}^T\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$$

for any symmetric positive definite matrix $\boldsymbol{A}$.

References

[1] H. Nickisch and M. Seeger. Convex variational Bayesian inference for large scale generalized linear models. In A. Danyluk, L. Bottou, and M. Littman, editors, *International Conference on Machine Learning 26*, volume 382, pages 761–768. ACM, 2009.

[2] M. Seeger. Low rank updates for the Cholesky decomposition. Technical report, University of California at Berkeley, 2004. See `lapmal.epfl.ch/papers/index.shtml`.

[3] M. Seeger. Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, 9:759–813, 2008.

[4] M. Seeger and H. Nickisch. Large scale Bayesian inference and experimental design for sparse linear models. *SIAM Journal of Imaging Sciences*, 4(1):166–199, 2011.