The Abdus Salam
International Centre
for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

2361-8

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 - Solutions to Exercises

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 — Solutions to Exercises

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

(b): If $\pi_j = \gamma_j^{-1}$, then

$$\frac{\partial - 2\log Z_Q}{\partial \pi_j} = \frac{-2}{Z_Q} \int P(\boldsymbol{y}|\boldsymbol{u}) \left(\prod_{i=1}^{q} e^{-\frac{1}{2} s_i^2/\gamma_i} \right) (-s_j^2/2)\, d\boldsymbol{u} = \mathrm{E}_Q[s_j^2].$$

(c):

$$\gamma_j \leftarrow \tau^{-1} \sqrt{\mathrm{E}_Q[s_j^2]}$$

(d): The equation in (c) is really coupled, since $\mathrm{E}_Q[s_j^2]$ depends on γ_j as well, but appears on the right hand side only. The complete minimization can be done by iterating the fixed point equation. The marginal $Q(s_j|\boldsymbol{y})$ does not have to be recomputed during this minimization, since we can always write

$$Q(s_j|\boldsymbol{y})' \propto Q(s_j|\boldsymbol{y}) e^{-\frac{1}{2}(\Delta\pi_j) s_j^2}, \quad \Delta\pi_j = \pi_j' - \pi_j.$$

2 Gaussian KL Minimization and Super-Gaussian Bounding

Obviously,

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

implies the statement. By definition of super-Gaussianity,

$$-\log t_j(s_j) \le s_j^2/(2\gamma_{*j}) + h_j(\gamma_{*j})/2,$$

so that

$$2\mathrm{E}_Q \left[-\log t_j(s_j) - s_j^2/(2\gamma_{*j}) \right] \le h_j(\gamma_{*j}).$$

3 Efficient Parameterization of Gaussian KL Minimization

(a) Immediate, given that

$$\nu_j = 2\mathrm{E}_Q[-\log t_j(s_j)]$$

and $\mathrm{E}_Q[\boldsymbol{u}\boldsymbol{u}^T] = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T$.

(b):

$$\nabla_{\boldsymbol{\Sigma}}\phi = -\boldsymbol{\Sigma}^{-1} + \boldsymbol{E} + \sum_{j=1}^{q} \pi_j \boldsymbol{b}_j \boldsymbol{b}_j^T.$$

Setting this equal to zero:

$$\boldsymbol{\Sigma}_*^{-1} = \boldsymbol{E} + \boldsymbol{B}^T(\operatorname{diag}\boldsymbol{\pi})\boldsymbol{B}.$$

4 Coordinate Update Algorithm for Gaussian KL Minimization

(a): Using the hint:

$$\Delta\pi_j = \frac{1}{\rho'_j} - \frac{1}{\rho_j}.$$

(b):

$$-p'_j + \boldsymbol{E}_{jj} + \frac{\partial \nu'_j}{\partial \rho'_j} = 0.$$

(c): This is just (a) in reverse. First, (b) implies that

$$\pi'_j = \frac{\partial \nu'_j}{\partial \rho'_j}.$$

Then,

$$\frac{1}{\rho'_j} = \Delta\pi_j + \frac{1}{\rho_j} \quad \Rightarrow \quad \rho'_j = \frac{\rho_j}{1 + (\Delta\pi_j)\rho_j}.$$

5 Spectral Analysis of Conjugate Gradients Algorithm

First,

$$P(\boldsymbol{A}) = P(\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T) = \sum_{j=0}^{k-1} \alpha_j (\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T)^j = \sum_{j=0}^{k-1} \alpha_j \boldsymbol{Q}\boldsymbol{\Lambda}^j\boldsymbol{Q}^T = \boldsymbol{Q}P(\boldsymbol{\Lambda})\boldsymbol{Q}^T,$$

because $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$. Then,

$$q(\boldsymbol{x}) = (1/2)\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{x} = (1/2)\boldsymbol{y}^T\boldsymbol{\Lambda}\boldsymbol{y} - \bar{\boldsymbol{b}}^T\boldsymbol{y}.$$

Since $\boldsymbol{A}^{-1} = \boldsymbol{Q}\boldsymbol{\Lambda}^{-1}\boldsymbol{Q}^T$, we have that $q_* = -(1/2)\boldsymbol{b}^T\boldsymbol{Q}\boldsymbol{\Lambda}^{-1}\boldsymbol{Q}^T\boldsymbol{b} = -(1/2)\bar{\boldsymbol{b}}^T\boldsymbol{\Lambda}^{-1}\bar{\boldsymbol{b}}$.

Next, $\boldsymbol{x}_k \in \mathcal{K}_k$, which is spanned by $\boldsymbol{A}^j \boldsymbol{b}$ for $j < k$. This means that $\boldsymbol{x}_k = P_k(\boldsymbol{A})\boldsymbol{b}$ for some polynomial with $\deg(P_k) < k$. Therefore,

$$\boldsymbol{y}_k = \boldsymbol{Q}^T P_k(\boldsymbol{A})\boldsymbol{b} = P_k(\boldsymbol{\Lambda})\boldsymbol{Q}^T\boldsymbol{b} = P_k(\boldsymbol{\Lambda})\bar{\boldsymbol{b}}.$$

Using the solutions from above,

$$\begin{aligned} q(\boldsymbol{x}_k) - q_* &= (1/2)\min_{P_k} \sum_{i=1}^n (\lambda_i y_{k,i}^2 - \bar{b}_i y_{k,i} + \bar{b}_i^2/\lambda_i) = (1/2)\min_{P_k}\sum_{i=1}^n \bar{b}_i^2(\lambda_i P_k(\lambda_i)^2 - P_k(\lambda_i) + 1/\lambda_i) \\ &= (1/2)\min_{P_k}\sum_{i=1}^n (\bar{b}_i^2/\lambda_i)(\lambda_i P_k(\lambda_i) - 1)^2. \end{aligned}$$

Let $Q_k(t) := tP_k(t) - 1$. As P_k runs over polynomials of degree $< k$, Q_k runs over polynomials of degree $\leq k$ with $Q_k(0) = -1$. The bound is nondecreasing in the $Q_k(\lambda_i)^2$ (which is why we can also use $-Q_k$ for the argument). Without assumptions on $\boldsymbol{b}$, we have to strive for small $|Q_k(\lambda_i)|$, especially for the smaller λ_i.

If $\{\lambda_1, \dots, \lambda_n\} = \{\kappa_1, \dots, \kappa_k\}$, pick the polynomial $Q_k(t) = [\prod_{j=1}^k (t - \kappa_j)]/[\prod_j \kappa_j]$. Here, $\prod_j \kappa_j > 0$ because all $\kappa_j > 0$ ($\boldsymbol{A}$ is positive definite). Then, $|Q_k(0)| = 1$ and $Q_k(\lambda_i) = 0$ for all i, so that $q(\boldsymbol{x}_k) = q_*$. Since q is strictly convex, it has a unique minimum, so $\boldsymbol{x}_k = \boldsymbol{x}_*$.

6 Super-Gaussian Bounding for Bernoulli Potentials

(a):

$$\frac{1}{1 + e^{-ys}} = \frac{e^{ys/2}}{e^{ys/2} + e^{-ys/2}},$$

so that

$$b = y/2, \quad \tilde{t}(s) = \frac{1}{2\cosh(bs)}.$$

$\tilde{t}(s)$ is even. Using the hint, $\log \tilde{t}(s) = -\log\cosh(bs) - \log 2$ is a convex function of $x = s^2$. This means that $\tilde{t}(s)$, and therefore $t(s)$, are super-Gaussian.

(b): It suffices to show that $f(x) = \log(1 + e^x)$ is convex.

$$f'(x) = \frac{e^x}{1 + e^x} = \frac{1}{1 + e^{-x}}, \quad f''(x) = \frac{e^{-x}}{(1 + e^{-x})^2} = f'(x)f'(-x) > 0,$$

so that $f(x)$ is strictly convex. This means that $t(s)$ is log-concave, and so is $\tilde{t}(s)$, since $\log \tilde{t}(s) = \log t(s) - bs$.

(c): We follow the argument given in the course. The MAP problem for the setup here would be

$$\min_{\boldsymbol{u}_*} \sigma^{-2}\|\boldsymbol{u}_*\|^2 - 2\sum_{j=1}^q \log t(s_{*j}), \quad \boldsymbol{s}_* = \boldsymbol{B}\boldsymbol{u}_*.$$

Now, $t(s_{*j}) = e^{bs_{*j}}\tilde{t}(s_{*j})$. In the course, we showed that for an even potential, the IL problem differs from MAP in that $\tilde{t}(s_{*j})$ is replaced by $\tilde{t}((z_j + s_{*j})^{1/2})$. Here,

$$\log t(s_{*j}) = bs_{*j} + \log \tilde{t}(s_{*j}) \to bs_{*j} + \log \tilde{t}((z_j + s_{*j}^2)^{1/2}),$$

so the IL problem is

$$\min_{\boldsymbol{u}_*} \sigma^{-2}\|\boldsymbol{u}_*\|^2 - 2\sum_{j=1}^{q}\left(bs_{*j} + \log\tilde{t}((z_j + s_{*j}^2)^{1/2})\right).$$

From (b), $\log\tilde{t}(s_{*j})$ is log-concave, which implies the convexity of $h(\gamma_j)$ and therefore of $\tilde{t}((z_j + s_{*j}^2)^{1/2})$ (this result is quoted in the course and proved in [1]), and bs_{*j} is linear, therefore convex.

7 Proximal Map for Inner Loop Optimization Problem

(a): If

$$s' = \mathrm{prox}(r) = \operatorname*{argmin}_s f(s;r),$$

then $f(s;-r) = f(-s;r)$, so that $\mathrm{prox}(-r) = -\mathrm{prox}(r)$. If $r > 0$ and $s < 0$, then $f(-s;r) < f(s;r)$, since $(-s-r)^2 < (s-r)^2$. Therefore, $\mathrm{prox}(r) \geq 0$.

(b): Recall that $s \geq 0$. Define $y = (1+s^2)^{1/2}$, so that

$$f(s;r) = \kappa y + \frac{1}{2}(s-r)^2.$$

The stationary equation is $df/ds = 0$:

$$\frac{\kappa s'}{y'} = r - s'.$$

$s' \leq r$ follows from $s'/y' \geq 0$. Also, $r > 0$ implies $s' > 0$. Moreover, $r = s'(1+\kappa/y') < s'(1+\kappa)$, since $y' > 1$, so that $s' > r/(1+\kappa)$. Finally, $y' > |s'|$, so that $s'/y' < 1$ and $r - s' < \kappa$. These inequalities can be used to bracket a solution for s'.

(c): Squaring both sides of the stationary equation gives

$$\kappa^2 s^2 = (1+s^2)(r-s)^2 \quad\Leftrightarrow\quad s^4 - 2rs^3 + (r^2+1-\kappa^2)s^2 - 2rs + r^2 = 0.$$

8 Bound on Marginal Variances

First,

$$\boldsymbol{A} = \sigma^{-2}\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{B}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{B}.$$

We have that $\mathrm{Var}_Q[s_j|\boldsymbol{y}] = \boldsymbol{b}_j^T\boldsymbol{A}^{-1}\boldsymbol{b}_j$, where $\boldsymbol{b}_j = \boldsymbol{B}^T\boldsymbol{\delta}_j$ is the j-th row of $\boldsymbol{B}$. Then,

$$\begin{aligned}\boldsymbol{b}_j^T\boldsymbol{A}^{-1}\boldsymbol{b}_j &= \max_{\boldsymbol{x}} 2\boldsymbol{b}_j^T\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \max_{\boldsymbol{x}} 2\boldsymbol{\delta}_j^T\boldsymbol{B}\boldsymbol{x} - \sigma^{-2}\|\boldsymbol{X}\boldsymbol{x}\|^2 - (\boldsymbol{B}\boldsymbol{x})^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{B}\boldsymbol{x}) \\ &\leq \max_{\boldsymbol{x}} 2\boldsymbol{\delta}_j^T(\boldsymbol{B}\boldsymbol{x}) - (\boldsymbol{B}\boldsymbol{x})^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{B}\boldsymbol{x}) \leq \max_{\boldsymbol{w}} 2\boldsymbol{\delta}_j^T\boldsymbol{w} - \boldsymbol{w}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{w} = \boldsymbol{\delta}_j^T\boldsymbol{\Gamma}\boldsymbol{\delta}_j = \gamma_j.\end{aligned}$$

The first $\leq$ is due to $\|\boldsymbol{X}\boldsymbol{x}\|^2 \geq 0$, the second due to the fact that $\boldsymbol{B}\boldsymbol{x}$ runs over a subspace of $\boldsymbol{w} \in \mathbb{R}^q$. The first and last $=$ are applications of the identity provided in the hint.

References

[1] 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.