Supplementary material

November 3, 2010

We provide small corrections to the main submission in this supplementary material to save the time of the reviewers (and apologize for the typos). We also provide optional additional details for some derivations to make the manuscript more easily verifiable.

1 Typos

- The paragraph on p.6 just above the section 4.3 should have "loss-insensitive fixed point $\mu_{q_{\text{sp}}^{\text{KL}}}$" instead of "non loss-sensitive fixed point $\mu_{q_{\text{sp}}^{\text{opt}}}$" – i.e. the updates converge to the min-KL solution rather than something which would be loss-calibrated such as $\mu_{q_{\text{sp}}^{\text{opt}}}$.
- Equation (23) should be corrected to:

$$p(y|x,\theta) = \Phi\left(\frac{y\, K_{x\mathcal{D}}\theta}{\tilde{\sigma}_x}\right)$$

- Equation (24) should be corrected to:

$$p_q(y|x) = \Phi\left(\frac{y\, K_{x\mathcal{D}}\mu_q}{\tilde{\sigma}_q(x)}\right)$$

- Equation (28) should have $\forall y \in \mathcal{Y}$ as well in the constraint.

2 Derivations notes

2.1 $\mu_{q_{\text{sp}}^{\text{KL}}}$ (in text left of p.6)

We want to find the μ_q which minimizes the KL expression given in (21) subject to the sparsity constraint $(\mu_q)_i = 0$ for $i \in T$. Writing $\tilde{\Lambda} \doteq \Sigma_{p_{\mathcal{D}}}^{-1}$ and setting the derivative to zero, we get that the non-zero components of μ_q (on the set S) are given by:

$$\mu_{q_{\text{sp}}^{\text{KL}}} = \tilde{\Lambda}_{SS}^{-1}\tilde{\Lambda}_{S\mathcal{D}}\mu_{p_{\mathcal{D}}}. \tag{2.1}$$

Substituting $\Sigma_{p_\mathcal{D}}^{-1} = K_{\mathcal{D}\mathcal{D}} + \sigma^{-2} K_{\mathcal{D}\mathcal{D}}^2$ and $\mu_{p_\mathcal{D}} = (K_{\mathcal{D}\mathcal{D}} + \sigma^2 I)^{-1}\mathbf{y}$, we have that:

$$\tilde{\Lambda}_{S\mathcal{D}}\mu_{p_\mathcal{D}} = K_{S\mathcal{D}}(\mathcal{I} + \sigma^{-2}K_{\mathcal{D}\mathcal{D}})(K_{\mathcal{D}\mathcal{D}} + \sigma^2 I)^{-1}\mathbf{y} = \sigma^{-2}K_{S\mathcal{D}}\mathbf{y}, \tag{2.2}$$

which is the convenient cancellation which enables us to avoid the inversion of the $N \times N$ matrix $K_{\mathcal{D}\mathcal{D}}$ which was previously needed to compute $\mu_{p_\mathcal{D}}$. Substituting (2.2) into (2.1) and expanding $\tilde{\Lambda}_{SS}$, we get

$$\mu_{q_{\text{sp}}^{\text{KL}}} = \left(\sigma^2 K_{SS} + K_{S\mathcal{D}}K_{\mathcal{D}S}\right)^{-1} K_{S\mathcal{D}}\,\mathbf{y},$$

which was the expression in the text. Note that similarly $\mu_{q_{\text{sp}}^{\text{opt}}} = \Lambda_{SS}^{-1}\Lambda_{S\mathcal{D}}\mu_{p_\mathcal{D}}$, but in this case $\Lambda_{S\mathcal{D}}\mu_{p_\mathcal{D}}$ doesn't simplify, hence we still need the $O(N^3)$ computation unfortunately in this case.

2.2 Linearized loss-EM update for GP regression (last paragraph of section 4.2)

We derive here the linearized loss-EM update (see table 2) with sparse constraints for GP regression which was claimed to be efficiently computable in $O(k^3 + Nk^2)$ in the last paragraph of section 4.2. As we mentioned in the text, $h_q(x) = \mu_q(x) = K_{x\mathcal{D}}\mu_q$ and so the linearized M-step is trivially $h^{t+1}(x) = K_{x\mathcal{D}}\mu_{q^{t+1}}$. We now look at the linearized E-step:

$$q^{t+1} = \arg\min_{q\in\mathcal{Q}} \;\; KL\left(q \| p_\mathcal{D}\right) + \frac{\mathcal{R}_q(h^t)}{M}. \tag{2.3}$$

Because h^t only depends on μ_{q^t} (and not Σ_{q^t}), we only need to optimize the RHS of (2.3) with respect to μ_q. So only keeping the terms depending on μ_q in (2.3) and using equation (19) to get an expression for $\mathcal{R}_q(h^t)$ (replace h_q with h^t and $p_\mathcal{D}$ with q) and equation (21) for the KL, we need to optimize:

$$\mu_q^\top \frac{\tilde{\Lambda}}{2}\mu_q - \mu_q^\top \tilde{\Lambda}\mu_{p_\mathcal{D}} + \mu_q^\top \frac{\Lambda}{M}\mu_q - 2\mu_q^\top \Lambda \mu_{q^t}.$$

Now letting μ_S be the non-zero coefficients of μ_q and setting the rest to zero, we get:

$$\mu_S^\top \left(\frac{\tilde{\Lambda}_{SS}}{2} + \frac{\Lambda_{SS}}{M}\right)\mu_S - \mu_S^\top \left(\tilde{\Lambda}_{S\mathcal{D}}\mu_{p_\mathcal{D}} + 2\frac{\Lambda_{S\mathcal{D}}}{M}\mu_{q^t}\right).$$

Setting the derivative to zero and solving for μ_S gives:

$$\mu_S^{t+1} = \breve{\Lambda}_{SS}^{-1}\left(\tilde{\Lambda}_{S\mathcal{D}}\mu_{p_\mathcal{D}} + 2\frac{\Lambda_{S\mathcal{D}}}{M}\mu_{q^t}\right), \tag{2.4}$$

where $\breve{\Lambda}_{SS} \doteq \tilde{\Lambda}_{SS} + 2\Lambda_{SS}/M$ is a $k \times k$ matrix which is computable in $O(Nk^2)$ time since $\tilde{\Lambda}_{SS} = K_{SS} + \sigma^{-2}K_{S\mathcal{D}}K_{\mathcal{D}S}$. Moreover, $\tilde{\Lambda}_{S\mathcal{D}}\mu_{p_\mathcal{D}}$ simplifies to $\sigma^{-2}K_{S\mathcal{D}}\mathbf{y}$ as we saw in equation (2.2) for the KL derivation. So the whole update can be

done in $O(k^3 + Nk^2)$ as we claimed. It is also loss-sensitive as it contains a $\Lambda_{S\mathcal{D}}$ term which has information about the test distribution $p(x)$. Unfortunately, this dependence is lost in the limit. First of all, we can see that this sequence of update does converge to a fixed point by noting that μ_S^t is multiplied by the matrix $A \doteq \breve{\Lambda}_{SS}^{-1}(2\Lambda_{SS}/M)$ which has norm smaller than one. We basically have $\mu_S^{t+1} = \mathbf{a} + A\mu_S^t$ and so μ_S^{t+1} is expressible with a geometric sum in A which converges as $t \to \infty$ because $||A|| < 1$. We can find the fixed point of (2.4) by putting $\mu_S^{t+1} = \mu_\infty$ on the LHS and $\mu_{q^t} = \mu_\infty$ on the RHS and solving for μ_∞ to get:

$$\mu_\infty = \tilde{\Lambda}_{SS}^{-1}\tilde{\Lambda}_{S\mathcal{D}}\mu_{p_\mathcal{D}}$$

and so $\mu_\infty = \mu_{q_{\text{sp}}^{\text{KL}}}$ by comparing with (2.1), as we said in the text.

2.3 Deriving the q-optimal action for GP classification (equation (25))

We note that the q-risk for the predictive setup becomes (by pushing the integral with respect to θ inside the sum terms of the generalization error $L(\theta, h)$):

$$\mathcal{R}_q(h) = \int_\mathcal{X} p(x) \left(\sum_{y \in \mathcal{Y}} p_q(y|x)\ell(y, h(x)) \right) dx.$$

So in our setup of section 4.3, we get $h_q(x)$ by minimizing the integrand pointwise:

$$h_q(x) = \underset{y' \in \{-1,+1\}}{\arg\min}\ \mathbb{I}_{\{y'=+1\}} c_+ \Phi\left(\frac{-K_{x\mathcal{D}}\mu_q}{\tilde{\sigma}_q(x)}\right) + \mathbb{I}_{\{y'=-1\}} c_- \Phi\left(\frac{K_{x\mathcal{D}}\mu_q}{\tilde{\sigma}_q(x)}\right).$$

So we want to choose $y' = +1$ when:

$$c_+ \Phi\left(\frac{-K_{x\mathcal{D}}\mu_q}{\tilde{\sigma}_q(x)}\right) < c_- \Phi\left(\frac{K_{x\mathcal{D}}\mu_q}{\tilde{\sigma}_q(x)}\right).$$

Using the fact that $\Phi(-a) = 1 - \Phi(a)$ and rearranging the terms give the choice function (25).