ELL 409 Minor

2.1 $$L(w) = \prod_{n=1}^{N} p(t_n | x_n; w) \quad \text{[IID assumption]}$$

$$= \prod_{n=1}^{N} \left[\frac{(w x_{n1} x_{n2})^{t_n} e^{-w x_{n1} x_{n2}}}{t_n!} \right].$$

2.2 Define $E(w) \triangleq -\log L(w)$

$$= -\sum_{n=1}^{N} \log \left[\frac{(w x_{n1} x_{n2})^{t_n} e^{-w x_{n1} x_{n2}}}{t_n!} \right]$$

$$= -\sum_{n=1}^{N} \left[t_n \log (w x_{n1} x_{n2}) - w x_{n1} x_{n2} - \log (t_n!) \right].$$

2.3 Set $\frac{\partial}{\partial w} E(w) = 0$

$$\frac{\partial}{\partial w} E(w) = \sum_{n=1}^{N} \left[\frac{-t_n}{w} + x_{n1} x_{n2} \right]$$

$$\Rightarrow \boxed{\hat{w}_{ML} = \frac{\sum_n t_n}{\sum_n x_{n1} x_{n2}}}.$$

2.4 Fraction of total no. of links between all district pairs to sum of population products of all pairs of districts; i.e., it is

the actual number of inter-district links for the whole data set, as a proportion of the total possible no. of such links. Hence it is the empirical probability of link formation between a random pair of individuals in this data set, which is indeed what w should represent, given that $w x_{n1} x_{n2}$ is the mean (expectation) of the Poisson for the total no. of links between a pair of districts.

2.5 $p(\underline{w} | X, \underline{t}) \propto p(\underline{t} | X; \underline{w}) \cdot p(\underline{w})$

(for discriminative modelling)

Here RHS $(\triangleq \tilde{\mathcal{L}}(w))$ $= \left[\prod_{n=1}^{N} \frac{(w x_{n1} x_{n2})^{t_n} e^{-w x_{n1} x_{n2}}}{t_n!} \right] \frac{\beta^{\alpha} w^{\alpha-1} e^{-\beta w}}{(\alpha-1)!}$.

2.6 Define $\tilde{E}(w) \triangleq -\log \tilde{\mathcal{L}}(w)$

$$= -\sum_{n=1}^{N} \left[t_n \log (w x_{n1} x_{n2}) - w x_{n1} x_{n2} - \log (t_n!) \right]$$

$$- \alpha \log \beta - (\alpha-1) \log w + \beta w + \log ((\alpha-1)!).$$

2.7 Set $\frac{\partial}{\partial w} \tilde{E}(w) = 0$

$$\frac{\partial}{\partial w} \tilde{E}(w) = \sum_{n=1}^{N} \left[\frac{-t_n}{w} + x_{n1} x_{n2} \right] - \frac{\alpha-1}{w} + \beta.$$

$$\Rightarrow \boxed{\hat{w}_{MAP} = \frac{\sum_n t_n + (\alpha - 1)}{\sum_n x_{n1} x_{n2} + \beta}}$$

2.8 In the absence of any data, we want $\frac{\alpha - 1}{\beta} = \frac{0.1}{100}$

$$\Rightarrow \underline{\beta = 1000\,(\alpha - 1)} \,.$$

2.9 Increasing α/β, while maintaining the above relationship between them, strengthens the prior.

3. → Bayesian perspective modulates the trade-off via the <u>prior</u> : a higher bias arises due to a more constrained or concentrated prior over hypothesis space, whereas a higher variance arises due to a more flexible or diffuse prior over hypothesis space.

→ So Bayesian models can control this trade-off via controlling the variance of the prior.

→ Fixing the prior hyperparameter(s)

which control its variance, e.g. α in the curve-fitting example discussed in class, is the way to find the sweet spot. As usual, such tuning could be done via validation of some form.

4.1 $$\tilde{E}(\underline{w}) = \frac{1}{2}\sum_n \underbrace{\left(y(x_n;\underline{w})}_{\triangleq y_n} - t_n\right)^2 + \frac{\lambda}{2}\|\underline{w}\|^2$$

$$\triangleq \sum_n \tilde{E}_n(\underline{w}) + \frac{\lambda}{2}\|\underline{w}\|^2$$

So $\tilde{E}_n(\underline{w}) = \frac{1}{2}(y_n - t_n)^2$ is the contribution of the n^{th} point to the error.

Now, for SGD:

$$\nabla_{\underline{w}}\left[\tilde{E}_n(\underline{w}) + \frac{\lambda}{2}\|\underline{w}\|^2\right]$$

$$= \nabla_{\underline{w}}\left[\frac{1}{2}(y_n - t_n)^2 + \frac{\lambda}{2}\|\underline{w}\|^2\right]$$

$$= \nabla_{\underline{w}}\left[\frac{1}{2}(y_n - t_n)^2 + \frac{\lambda}{2}(w_0^2 + w_1^2 + w_2^2)\right]$$

$$= \begin{pmatrix} (y_n - t_n)\frac{\partial y_n}{\partial w_0} + \lambda w_0 \\ (y_n - t_n)\frac{\partial y_n}{\partial w_1} + \lambda w_1 \\ (y_n - t_n)\frac{\partial y_n}{\partial w_2} + \lambda w_2 \end{pmatrix} = \begin{pmatrix} y_n - t_n + \lambda w_0 \\ (y_n - t_n)x_n + \lambda w_1 \\ (y_n - t_n)x_n^2 + \lambda w_2 \end{pmatrix}.$$

4.2 $$\underline{w}^{(\tau+1)} = \underline{w}^{(\tau)} - \eta \nabla_{\underline{w}} \tilde{E}(\underline{w})\Big|_{\underline{w}^{(\tau)}}$$

So here: (using SGD)

where (x_n, t_n) is the datapt. used at iteration $(\tau+1)$

$$\begin{cases} w_0^{(\tau+1)} = w_0^{(\tau)} - \eta \left[y(x_n; \underline{w}^{(\tau)}) - t_n + \lambda w_0^{(\tau)} \right] \\ w_1^{(\tau+1)} = w_1^{(\tau)} - \eta \left[(y(x_n; \underline{w}^{(\tau)}) - t_n) x_n + \lambda w_1^{(\tau)} \right] \\ w_2^{(\tau+1)} = w_2^{(\tau)} - \eta \left[(y(x_n; \underline{w}^{(\tau)}) - t_n) x_n^2 + \lambda w_2^{(\tau)} \right] \end{cases}$$

4.3 For each weight, L2 reg. is contributing a term of the form $\underline{-\eta \lambda w_i^{(\tau)}}$ to the update equation. Since $\eta > 0$ and $\lambda > 0$, this means, generally, a reduction in the magnitude of w_i. Note that both η and λ are typically small, so $\eta\lambda$ is expected to be $\ll 1$ in most cases. Hence we can also think of the update as $w_i^{(\tau+1)} = \underline{(1-\eta\lambda)} w_i^{(\tau)} + \ldots$ So it systematically tries to suppress the wts. on each iteration.