Appendix

Appendix A: Derivation of the Upper Bound

We provide details on deriving the variational bound of the expected hinge loss in (4). To simplify notations, we derive the bound for a single data point. For a data set with N examples, a simple summation will give the final bound. Define $g(\boldsymbol{\theta}; \mathbf{x}) := \mathbb{E}_p[\log \phi(y|\tilde{\mathbf{x}}, \boldsymbol{\theta})]$. We have

$$\begin{aligned} g(\boldsymbol{\theta}; \mathbf{x}) &= \mathbb{E}_p\Big[\log \int \frac{1}{\sqrt{2\pi\lambda}} \exp\Big\{-\frac{(\lambda + c\zeta))^2}{2\lambda}\Big\} d\lambda\Big] \\ &= \mathbb{E}_p\Big[\log \int \frac{q(\lambda)}{q(\lambda)\sqrt{2\pi\lambda}} \exp\Big\{-\frac{(\lambda + c\zeta)^2}{2\lambda}\Big\} d\lambda\Big] \\ &\geq \mathbb{E}_p\Big[\mathbb{E}_{q(\lambda)} \log \frac{1}{q(\lambda)\sqrt{2\pi\lambda}} \exp\Big\{-\frac{(\lambda + c\zeta)^2}{2\lambda}\Big\}\Big] \\ &= \Big\{H(\lambda) - \frac{1}{2}\mathbb{E}_q[\log \lambda] - \mathbb{E}_q\Big[\frac{1}{2\lambda}\mathbb{E}_p(\lambda + c\zeta)^2\Big]\Big\} + c' \end{aligned}$$

where λ is the augmented variable, and c' is a constant. Note that the data augmentation at the first two equalities are exact and does not incur any approximation. The approximation is from the assumption that $q(\lambda)$ is independent of the "corrupted" observations $\tilde{\mathbf{x}}$. If there is no uncertainty in the feature corruption (e.g., the corruption level in the dropout (or blankout) noise is 0), the bound is tight. That is, the optimal solution of q will give the original hinge loss.

Appendix B. Proof of Lemma 1

Proof. Ignore the ℓ_2-norm regularizer, we have the objective of the M-step:

$$\mathcal{L}_{[\mathbf{w}]} = \sum_{n=1}^{N} \mathbb{E}_p\left[c\zeta_n + \frac{c^2}{2}\gamma_n\zeta_n^2\right], \tag{23}$$

where $\gamma_n := \mathbb{E}_q[\lambda_n^{-1}]$. Using the definition of $\zeta_n := \ell - y_n\mathbf{w}^\top\tilde{\mathbf{x}}_n$ and ignoring the constants, we have the simplified objective function (again without the ℓ_2-regularizer):

$$\begin{aligned} \mathcal{L}_{[\mathbf{w}]} &= \sum_{n=1}^{N} \mathbb{E}_p\left[\frac{c^2}{2}\gamma_n\mathbf{w}^\top\tilde{\mathbf{x}}_n\tilde{\mathbf{x}}_n^\top\mathbf{w} - (c + \ell c^2\gamma_n)y_n\mathbf{w}^\top\tilde{\mathbf{x}}_n\right] \\ &= \frac{c^2}{2}\sum_{n=1}^{N}\gamma_n\mathbb{E}_p\left[\mathbf{w}^\top\tilde{\mathbf{x}}_n\tilde{\mathbf{x}}_n^\top\mathbf{w} - 2y_n^h\mathbf{w}^\top\tilde{\mathbf{x}}_n\right] \\ &= \frac{c^2}{2}\sum_{n=1}^{N}\gamma_n\mathbb{E}_p\left[(\mathbf{w}^\top\tilde{\mathbf{x}}_n - y_n^h)^2\right], \end{aligned} \tag{24}$$

where $y_n^h := (\frac{1}{c\gamma_n} + \ell)y_n$ is the re-weighted label.

We now derive the equations to compute γ_n. Let x be a random variable and $y = f(x)$ is a function of x. Then, we have the transformation rule of probability distributions, $p(x) = p(f(x))|\frac{df(x)}{dx}|$. For our case, let $x = \lambda_n$, and $f(x) = \frac{1}{\lambda_n}$, we have $q(\lambda_n) = \frac{1}{\lambda_n^2}q(\frac{1}{\lambda_n})$. Then

$$\begin{aligned} \mathbb{E}_{q(\lambda_n)}[\lambda_n^{-1}] &= \int_0^\infty q(\lambda_n)\frac{1}{\lambda_n}d\lambda_n \\ &= \int_0^\infty q\left(\frac{1}{\lambda_n}\right)\frac{1}{\lambda_n^3}d\lambda_n \\ &= \int_\infty^0 q(\mu_n)\mu_n^3 d\mu_n^{-1} \ \ (\text{define } \mu_n = \frac{1}{\lambda_n}) \\ &= \int_0^\infty q(\mu_n)\mu_n d\mu_n \\ &= \mathbb{E}_{q(\lambda_n^{-1})}[\lambda_n^{-1}]. \end{aligned} \tag{25}$$

Since $q(\lambda_n^{-1})$ is an inverse Gaussian distribution as shown in Eq. (11), it is easy to get

$$\mathbb{E}_{q(\lambda_n)}[\lambda_n^{-1}] = \mathbb{E}_{q(\lambda_n^{-1})}[\lambda_n^{-1}] = \frac{1}{c\sqrt{\mathbb{E}[\zeta_n^2]}}. \tag{26}$$

Combining the above results finishes the proof of Lemma 1. □