Uncertainty-Aware Pseudo-labeling for Quantum Calculations (Supplementary Material)

Kexin Huang[1,2] **Vishnu Sresht**[1] **Brajesh Rai**[1] **Mykola Bordyuh**[1]

[1]Machine Learning and Computational Sciences, Pfizer Inc., Cambridge, Massachusetts, USA
[2]Department of Computer Science, Stanford University, Stanford, California, USA

1 ENTROPY OF STUDENT'S T-DISTRIBUTION

While the entropy of the Student's t-distribution is well known, we derive it for completeness. Student's probability distribution defined in terms of location γ, scale factor σ_{st}^2 and ν_{st} degrees of freedom is

$$\begin{aligned} p(y;\gamma,\sigma_{st}^2,\nu_{st}) &= \mathrm{St}(y;\gamma,\sigma_{st}^2,\nu_{st}) \\ &= \frac{\Gamma(\frac{\nu_{st}+1}{2})}{\sqrt{\nu_{st}\,\pi\sigma_{st}^2}\,\Gamma(\frac{\nu_{st}}{2})}\left(1+\frac{1}{\nu_{st}}\frac{(y-\gamma)^2}{\sigma_{st}^2}\right)^{-\frac{\nu_{st}+1}{2}}, \end{aligned} \tag{1}$$

where Γ i a gamma function. Student's t-distribution can be written in terms of beta function $\mathrm{B} = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ if we take advantage of the fact that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

$$\begin{aligned} \mathrm{p}(y;\sigma_{st}^2,\nu_{st}) &= \mathrm{St}(y;\sigma_{st},\nu_{st}) \\ &= \frac{1}{\sqrt{\nu_{st}\,\sigma_{st}^2}\,\mathrm{B}(\frac{1}{2},\frac{\nu_{st}}{2})}\left(1+\frac{1}{\nu_{st}}\frac{(y-\gamma)^2}{\sigma_{st}^2}\right)^{-\frac{\nu_{st}+1}{2}}. \end{aligned} \tag{2}$$

In the main text of the manuscript, we used empirical estimate of Student's t-distribution, which corresponds to evaluation at the highest mode $y=\gamma$. Empirical estimation of probability becomes:

$$\begin{aligned} \mathrm{p}^{\mathrm{emp}}(y=\gamma;\sigma_{st}^2,\nu_{st}) &\approx \mathrm{St}(y=\gamma;\sigma_{st},\nu_{st}) \\ &= \frac{1}{\sqrt{\nu_{st}\,\sigma_{st}^2}\,\mathrm{B}(\frac{1}{2},\frac{\nu_{st}}{2})}. \end{aligned} \tag{3}$$

If we introduce a new variable $t = \frac{y-\gamma}{\sigma_{st}}$, Student's t-distribution converts into the standard form with probability density

$$\mathrm{p}(t;\nu_{st}) = \mathrm{St}(t;\nu_{st}) = \frac{1}{\sqrt{\nu_{st}}\,\mathrm{B}(\frac{1}{2},\frac{\nu_{st}}{2})}\left(1+\frac{t^2}{\nu_{st}}\right)^{-\frac{\nu_{st}+1}{2}}. \tag{4}$$

1.1 PROPOSITION

Proposition: Entropy of the generalized and standard Student's t-distributions are related via the formula

$$\mathcal{H}(y;\sigma_{st}^2,\nu_{st}) = \mathcal{H}(t;\nu_{st}) + \frac{1}{2}\log\sigma_{st}^2\,. \tag{5}$$

Proof: The transformation $t = \mathrm{g}(y) = \frac{y-\gamma}{\sigma_{st}}$ is bijective and invertible with the inverse transformation $y = g^{-1}(t) = \sigma_{st}\,t+\gamma$. The Jacobin of the transformation g is $\frac{\mathrm{d}}{\mathrm{d}y}\,g(y) = \frac{1}{\sigma_{st}}$. According to the change of variables probability density formula

$$\mathrm{p}_y\,(y;\sigma_{st}^2,\nu_{st}) = \mathrm{p}_t\,(g(y);\nu_{st})\,|\,\frac{\mathrm{d}}{\mathrm{d}y}\,g(y)\,|. \tag{6}$$

The equation for the entropy transformation (equation 5) follows directly from the definition of the entropy.

To find the generalized entropy, we just need to calculate the entropy of the standard Student's t-distribution

$$\begin{aligned} \mathcal{H}(t;\nu_{st}) = &-\int_{-\infty}^{+\infty}\mathrm{p}(t;\nu_{st})\log\mathrm{p}(t;\nu_{st})\,\mathrm{d}t \\ &= \log\left(\sqrt{\nu_{st}}\,\mathrm{B}(\frac{1}{2},\frac{\nu_{st}}{2})\right)\int_{-\infty}^{+\infty}\mathrm{p}(t;\nu_{st})\,\mathrm{d}t \\ &+\frac{\nu_{st}+1}{2}\int_{-\infty}^{+\infty}\log(1+\frac{t^2}{\nu_{st}})\,\mathrm{p}(t;\nu_{st})\,\mathrm{d}t \\ &+\log\left(\sqrt{\nu_{st}}\,\mathrm{B}(\frac{1}{2},\frac{\nu_{st}}{2})\right) \\ &+\frac{\nu_{st}+1}{2}\int_{-\infty}^{+\infty}\log(1+\frac{t^2}{\nu_{st}})\,\mathrm{p}(t;\nu_{st})\,\mathrm{d}t\,(*). \end{aligned} \tag{7}$$

To find the second integral, we make a substitution $x = \frac{t^2}{\nu_{st}}$

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

and obtain

$$
\begin{aligned}
(*) &= \frac{\nu_{st}+1}{2\,\mathrm{B}(\frac{1}{2},\frac{\nu_{st}}{2})} \int_0^{+\infty} \log(1+x)\,(1+x)^{-\frac{\nu_{st}+1}{2}} \frac{\mathrm{d}x}{\sqrt{x}} \\
&= -\frac{\nu_{st}+1}{\mathrm{B}(\frac{1}{2},\frac{\nu_{st}}{2})} \frac{\partial}{\partial \nu_{st}} \int_0^{+\infty} (1+x)^{-\frac{\nu_{st}+1}{2}} \frac{\mathrm{d}x}{\sqrt{x}} \\
&= -\frac{\nu_{st}+1}{\mathrm{B}(\frac{1}{2},\frac{\nu_{st}}{2})} \frac{\partial}{\partial \nu_{st}} \int_0^{1} x^{\frac{\nu_{st}}{2}-1}\,(1-x)^{\frac{1}{2}-1}\,\mathrm{d}x \\
&= -\frac{\nu_{st}+1}{\mathrm{B}(\frac{1}{2},\frac{\nu_{st}}{2})} \frac{\partial}{\partial \nu_{st}} \mathrm{B}(\frac{1}{2},\frac{\nu_{st}}{2}) \\
&= -(\nu_{st}+1)\,\frac{\partial}{\partial \nu_{st}} \log \mathrm{B}(\frac{1}{2},\frac{\nu_{st}}{2}) \\
&= -(\nu_{st}+1)\,\frac{\partial}{\partial \nu_{st}} \left(\log\Gamma(\frac{\nu_{st}}{2}) - \log\Gamma(\frac{\nu_{st}+1}{2}) \right) \\
&= \frac{\nu_{st}+1}{2} \left(\Psi(\frac{\nu_{st}+1}{2}) - \Psi(\frac{\nu_{st}}{2}) \right),
\end{aligned}
\tag{8}
$$

where digamma function is defined as $\Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$. Putting all the terms together, the entropy of the standard Student's t-distribution becomes

$$
\begin{aligned}
\mathcal{H}(t;\nu_{st}) = \frac{\nu_{st}+1}{2} &\left(\Psi(\frac{\nu_{st}+1}{2}) - \Psi(\frac{\nu_{st}}{2}) \right) \\
&+ \log\left(\sqrt{\nu_{st}}\,\mathrm{B}(\frac{1}{2},\frac{\nu_{st}}{2}) \right).
\end{aligned}
\tag{9}
$$

The final formula for the entropy of the labels y is given by

$$
\begin{aligned}
\mathcal{H}(y;\sigma^2_{st},\nu_{st}) = \frac{\nu_{st}+1}{2} &\left(\Psi(\frac{\nu_{st}+1}{2}) - \Psi(\frac{\nu_{st}}{2}) \right) \\
&+ \log\left(\sqrt{\nu_{st}}\,\mathrm{B}(\frac{1}{2},\frac{\nu_{st}}{2}) \right) + \frac{1}{2}\,\log\,\sigma^2_{st}\,.
\end{aligned}
\tag{10}
$$

2 PSEUDO-LABELING, ENTROPY MINIMIZATION AND ALEATORIC UNCERTAINTY

In the main section of the text, we have considered the case where observed targets $(y_1, \cdots, y_i) \sim \mathcal{N}(\mu, \sigma^2)$ are drawn from the Normal distribution with unknown mean and variance μ and σ^2 and we have imposed a prior on them. The problem is significantly simplified if we treat μ and σ^2 in a deterministic way, such that our model f outputs two parameters μ and σ^2. The model here is able to estimate aleatoric (data) uncertainty σ^2 but unable to model epistemic (model) uncertainty. By minimizing negative log-likelihood, the loss is significantly simpler than in Eq. 4 in the main text.

$$
\mathcal{L}_i = -\log \mathcal{N}(y_i;\mu_i,\sigma_i^2) = \frac{2\pi\sigma_i^2}{2} + \frac{(y_i-\mu_i)^2}{2\sigma_i^2}. \tag{11}
$$

Empirical estimate of the entropy on the unlabeled data set becomes

$$
\begin{aligned}
\mathcal{H}(\mathcal{Y}\,|\,\mathcal{U}) &= \sum_{\mathbf{x}_i \in \mathcal{U}} \mathrm{E}_{y\sim \mathrm{p}_\theta(y\,|\,\mathbf{x}_i)}[-\log \mathrm{p}_\theta(y\,|\,\mathbf{x}_i)] \\
&\approx -\sum_{\mathbf{x}_i \in \mathcal{U}} \mathcal{E}_i^{emp}\,[\log \mathrm{p}_\theta(y\,|\,\mathbf{x}_i)]
\end{aligned}
\tag{12}
$$

with log probability weights $\mathcal{E}_i^{emp} = \frac{1}{\sqrt{2\pi\sigma_i^2}}$. One can notice that the weights are inversely related to aleatoric uncertainties $\mathcal{E}_i \sim (\sigma_i^2)^{-\frac{1}{2}}$.