From Softmax to Dirichlet: Evidential Learning for Semi-supervised Semantic Segmentation

Supplementary Material

In this supplementary material, we provide the detailed mathematical derivations omitted from the main paper. Sec. A demonstrates the derivation of the Dirichlet posterior distribution via Bayesian inference (referencing Sec. 3.2), and Sec. B details the derivation of the proposed Evidential Cross-Entropy loss (l_{ce}^{dir}) (referencing Sec. 3.4).

A. Derivation of the Dirichlet Posterior

In this section, we provide the rigorous mathematical derivation for the posterior distribution of class probabilities, expanding on the concepts introduced in Sec. 3.2 of the main paper.

A.1. Dirichlet Prior and Multivariate Beta Function

We model the class probability vector $\boldsymbol{p} = [p_1, \ldots, p_K]$ (where $\sum p_i = 1$) using a Dirichlet distribution. The probability density function (PDF) is defined as:

$$Dir(\boldsymbol{p}; \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} p_i^{\alpha_i - 1} \tag{14}$$

Here, $B(\boldsymbol{\alpha})$ is the **Multivariate Beta function**, which serves as the normalizing constant ensuring the PDF integrates to 1 over the K-dimensional simplex. It is defined using the **Gamma function** $\Gamma(\cdot)$ as:

$$B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}, \tag{15}$$

where the Gamma function $\Gamma(z)$ is defined by the integral:

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt \tag{16}$$

In our method, we adopt a *non-informative uniform prior* by setting $\boldsymbol{\alpha}^{prior} = \mathbf{1} = [1, \ldots, 1]$. Consequently, the prior distribution simplifies to:

$$P(\boldsymbol{p}) = Dir(\boldsymbol{p}; \mathbf{1}) \propto \prod_{i=1}^{K} p_i^{1-1} = 1 \tag{17}$$

A.2. Generalized Likelihood Function

The network collects evidence $\boldsymbol{e} = [e_1, \ldots, e_K]$ from the data. In standard Bayesian inference for categorical data, the likelihood of observing discrete counts $\boldsymbol{e}$ given probabilities $\boldsymbol{p}$ follows a **Multinomial distribution**:

$$P(\boldsymbol{x}|\boldsymbol{p}) = \frac{(\sum_i e_i)!}{e_1! e_2! \ldots e_K!} \prod_{i=1}^{K} p_i^{e_i} \tag{18}$$

Mathematically, the factorial function is generalized to the domain of real numbers via the **Gamma function**, satisfying the property $n! = \Gamma(n+1)$. By applying this standard analytical continuation, the Multinomial likelihood is naturally extended to the continuous domain, accommodating *continuous evidence values* **e** (e.g., neural network activations) as follows:

$$P(\boldsymbol{x}|\boldsymbol{p}) = \frac{\Gamma(\sum_i e_i + 1)}{\prod_i \Gamma(e_i + 1)} \prod_{i=1}^{K} p_i^{e_i} \tag{19}$$

Since the term involving the Gamma functions depends solely on $\boldsymbol{e}$ (which is fixed for a given input sample) and is independent of $\boldsymbol{p}$, the likelihood can be simplified for the purpose of posterior derivation:

$$P(\boldsymbol{x}|\boldsymbol{p}) \propto \prod_{i=1}^{K} p_i^{e_i} \tag{20}$$

A.3. Posterior Derivation via Bayes' Theorem

Using Bayes' theorem, we compute the posterior distribution by combining the prior and the likelihood:

$$\begin{aligned} P(\boldsymbol{p}|\boldsymbol{x}) &= \frac{P(\boldsymbol{x}|\boldsymbol{p})P(\boldsymbol{p})}{P(\boldsymbol{x})} \\ &\propto \left(\prod_{i=1}^{K} p_i^{e_i}\right) \cdot \left(\prod_{i=1}^{K} p_i^{1-1}\right) \\ &\propto \prod_{i=1}^{K} p_i^{(e_i+1)-1} \end{aligned} \tag{21}$$

We observe that the functional form of the resulting kernel, $\prod p_i^{(e_i+1)-1}$, matches the definition of the Dirichlet PDF with updated concentration parameters $\boldsymbol{\alpha}^{posterior} = \boldsymbol{e} + \mathbf{1}$. Thus, the posterior distribution is:

$$\boldsymbol{p}|\boldsymbol{x} \sim Dir(\boldsymbol{p}; \mathbf{1} + \boldsymbol{e}) \tag{22}$$

B. Derivation of Loss Function ℓ_{ce}^{dir}

In Sec. 3.4, we propose minimizing the expected cross-entropy loss over the induced posterior distribution. Rewrite the objective function in Eq. (12):

$$\begin{aligned} \mathcal{L} &= \mathbb{E}_{\boldsymbol{p} \sim Dir(\boldsymbol{p};\boldsymbol{\alpha})} \left[-\sum_{i=1}^{K} y_i \log p_i \right] \\ &= -\sum_{i=1}^{K} y_i \cdot \mathbb{E}_{\boldsymbol{p} \sim Dir(\boldsymbol{p};\boldsymbol{\alpha})} [\log p_i] \end{aligned} \tag{23}$$

where $\boldsymbol{\alpha} = \boldsymbol{e} + \mathbf{1}$. To derive Eq. (13), we need to solve for the expectation $\mathbb{E}[\log p_j]$.

B.1. Integral Expansion

By definition, the expectation of $\log p_j$ under the Dirichlet distribution is the integral over the simplex $\mathcal{S}_K$:

$$\begin{aligned}\mathbb{E}[\log p_j] &= \int_{\mathcal{S}_K} \log p_j \cdot Dir(\boldsymbol{p}; \boldsymbol{\alpha}) d\boldsymbol{p} \\ &= \int_{\mathcal{S}_K} \log p_j \cdot \left(\frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} p_k^{\alpha_k - 1} \right) d\boldsymbol{p} \\ &= \frac{1}{B(\boldsymbol{\alpha})} \int_{\mathcal{S}_K} \log p_j \prod_{k=1}^{K} p_k^{\alpha_k - 1} d\boldsymbol{p}\end{aligned} \tag{24}$$

We recall that the Multivariate Beta function $B(\boldsymbol{\alpha})$ is defined as the integral:

$$B(\boldsymbol{\alpha}) = \int_{\mathcal{S}_K} \prod_{k=1}^{K} p_k^{\alpha_k - 1} d\boldsymbol{p} \tag{25}$$

We differentiate $B(\boldsymbol{\alpha})$ with respect to the parameter α_j:

$$\begin{aligned}\frac{\partial B(\boldsymbol{\alpha})}{\partial \alpha_j} &= \frac{\partial}{\partial \alpha_j} \int_{\mathcal{S}_K} \prod_{k=1}^{K} p_k^{\alpha_k - 1} d\boldsymbol{p} \\ &= \int_{\mathcal{S}_K} \frac{\partial (p_j^{\alpha_j - 1})}{\partial \alpha_j} \prod_{k \neq j} p_k^{\alpha_k - 1} d\boldsymbol{p} \\ &= \int_{\mathcal{S}_K} (\log p_j) p_j^{\alpha_j - 1} \prod_{k \neq j} p_k^{\alpha_k - 1} d\boldsymbol{p} \\ &= \int_{\mathcal{S}_K} \log p_j \prod_{k=1}^{K} p_k^{\alpha_k - 1} d\boldsymbol{p}\end{aligned} \tag{26}$$

Substituting this integral back into the expectation formula in Eq. (24), we obtain:

$$\mathbb{E}[\log p_j] = \frac{1}{B(\boldsymbol{\alpha})} \frac{\partial B(\boldsymbol{\alpha})}{\partial \alpha_j} = \frac{\partial \log B(\boldsymbol{\alpha})}{\partial \alpha_j} \tag{27}$$

B.2. Connection to the Digamma Function

We expand $\log B(\boldsymbol{\alpha})$ using the Gamma function definition provided in Eq. (16):

$$\log B(\boldsymbol{\alpha}) = \sum_{k=1}^{K} \log \Gamma(\alpha_k) - \log \Gamma \left(\sum_{k=1}^{K} \alpha_k \right) \tag{28}$$

We introduce the **Digamma function** $\psi(x)$, defined as the logarithmic derivative of the Gamma function:

$$\psi(x) = \frac{d}{dx} \log \Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)} \tag{29}$$

Differentiating $\log B(\boldsymbol{\alpha})$ with respect to α_j:

$$\begin{aligned}\frac{\partial \log B(\boldsymbol{\alpha})}{\partial \alpha_j} &= \psi(\alpha_j) - \psi \left(\sum_{k=1}^{K} \alpha_k \right) \cdot \frac{\partial (\sum \alpha_k)}{\partial \alpha_j} \\ &= \psi(\alpha_j) - \psi \left(\sum_{k=1}^{K} \alpha_k \right)\end{aligned} \tag{30}$$

B.3. Final Loss Formulation

Substituting the expectation result back into the original loss function $\mathcal{L}$:

$$\begin{aligned}\ell_{ce}^{dir} &= - \sum_{j=1}^{K} y_j \left(\psi(\alpha_j) - \psi \left(\sum_{i=1}^{K} \alpha_i \right) \right) \\ &= \sum_{j=1}^{K} y_j \left(\psi \left(\sum_{i=1}^{K} \alpha_i \right) - \psi(\alpha_j) \right)\end{aligned} \tag{31}$$

This derivation confirms Eq. (13) presented in the main paper.