A note of Bayesian inference on HDP-PCFG

Tomohiko Nakamura

February 28, 2013

1 Projected gradient on β estimation

The generative model of β was expressed as

$$\beta \sim \mathrm{GEM}(\beta; \alpha), \tag{1}$$

where $\mathrm{GEM}(\beta; \alpha)$ means sticking break process such as

$$\beta_i := \tilde{\beta}_i \prod_{j<i} (1 - \tilde{\beta}_j), \ \ \tilde{\beta}_i \sim \mathrm{Beta}(\tilde{\beta}_i; 1, \alpha). \tag{2}$$

Here, we apply projected gradient on β to the following objective function,

$$L(\beta) := \ln \mathrm{GEM}(\beta; \alpha) + \sum_{z=1}^{K} E_q \ln \mathrm{Dir}(\phi_z^B | \alpha^B \beta \beta^\top). \tag{3}$$

The update rule with step size η is given as

$$\beta \leftarrow L(\beta) + \eta \frac{\partial L}{\partial \beta}(\beta). \tag{4}$$

Preparation

Let $L_{\mathrm{prior}}(\beta)$ and $L_{\mathrm{rules}}(\beta)$ donote the first term and second one of (3). We must change variables $\tilde{\beta}$ to β to derive an update rule on β. The definition of probability density function gives us

$$\int_{[0,1]^{K-1}} \prod_{i=1}^{K-1} \Big(\mathrm{Beta}(\tilde{\beta}_i; 1, \alpha) \Big) d\tilde{\beta}_{1:K-1} = 1 \quad \Leftrightarrow \quad \int_{\mathcal{T}} \mathrm{GEM}(\beta; \alpha) d\beta_{1:K-1} = 1, \tag{5}$$

$$\mathcal{T} := \left\{ (\beta_1, \cdots, \beta_{K-1}); \forall z \in [1, K-1], \beta_z \geq 0 \text{ and } \sum_{z=1}^{K-1} \beta_z \leq 1 \right\}, \tag{6}$$

and then we derive $\mathrm{GEM}(\beta; \alpha)$ distribution as follows;

$$\begin{aligned} \int_{[0,1]^{K-1}} \prod_{i=1}^{K-1} \Big(\mathrm{Beta}(\tilde{\beta}_i; 1, \alpha) \Big) d\tilde{\beta}_{1:K-1} &= \int_{\mathcal{T}} \prod_{i=1}^{K-1} \Big(\mathrm{Beta}(\tilde{\beta}_i; 1, \alpha) \Big) \frac{d\tilde{\beta}}{d\beta} d\beta_{1:K-1}, & (7) \\ &= \int_{\mathcal{T}} \prod_{i=1}^{K-1} \Big(\mathrm{Beta}(\beta_i T_i^{-1}; 1, \alpha) \Big) |J(\beta)| d\beta_{1:K-1}, & (8) \end{aligned}$$

$$\Leftrightarrow \int_{\mathcal{T}} \mathrm{GEM}(\beta; \alpha) d\beta_{1:K-1} = \int_{\mathcal{T}} \left[\prod_{i=1}^{K-1} \Big(\mathrm{Beta}(\beta_i T_i^{-1}; 1, \alpha) \Big) |J(\beta)| \right] d\beta_{1:K-1}, \tag{9}$$

where

$$T_i := 1 - \sum_{j<i} \beta_j \Big(= \prod_{j<i} (1 - \tilde{\beta}_j) \Big), \tag{10}$$

$$J(\beta) := \frac{d\tilde{\beta}}{d\beta} = \begin{bmatrix} T_1^{-1} & 0 & \cdots & 0 \\ * & T_2^{-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ * & \cdots & * & T_{K-1}^{-1} \end{bmatrix}. \tag{11}$$

The probability density function $\mathrm{GEM}(\beta;\alpha)$ is

$$\prod_{i=1}^{K-1}\Big(\mathrm{Beta}(\beta_i T_i^{-1};1,\alpha)\Big)|J(\beta)| = \prod_{i=1}^{K-1}\Big(\mathrm{Beta}(\beta_i T_i^{-1};1,\alpha)T_i^{-1}\Big), \tag{12}$$

$$= \prod_{i=1}^{K-1}\Big(\frac{\Gamma(\alpha+1)}{\Gamma(\alpha)\Gamma(1)}(1-\beta_i T_i^{-1})^{\alpha-1}T_i^{-1}\Big), \tag{13}$$

because $|J(\beta)| = \prod_{i=1}^{K-1} T_i^{-1}$.

The derivation on $L_{\mathrm{prior}}(\beta)$

$$L_{\mathrm{prior}}(\beta) = \ln \mathrm{GEM}(\beta;\alpha), \tag{14}$$

$$=_c \sum_{i=1}^{K-1}\Big((\alpha-1)\ln(1-\beta_i T_i^{-1}) - \ln T_i\Big), \tag{15}$$

$$= \sum_{i=1}^{K-1}\Big((\alpha-1)\ln(T_{i+1}T_i^{-1}) - \ln T_i\Big), \tag{16}$$

$$= (\alpha-1)\Big(\ln T_K - \ln T_1\Big) - \sum_{i=1}^{K-1}\ln T_i, \tag{17}$$

$$= (\alpha-1)T_K - \sum_{i=1}^{K-1}\ln T_i, \tag{18}$$

because $T_1 = 1$, and $=_c$ expresses equal without constant values. The partial differential is derived that

$$\frac{\partial L_{\mathrm{prior}}(\beta)}{\partial \beta_k} = -\frac{\alpha-1}{T_K} - \sum_{i=k+1}^{K-1}\frac{1}{T_i}. \tag{19}$$

Note that $\partial T_z/\partial \beta_k = -\mathbb{I}[z>k]$.

The derivation on $L_{\mathrm{rules}}(\beta)$

$$L_{\mathrm{rules}}(\beta) = \sum_{z=1}^{K} E_{q(\phi)}[\ln \mathrm{Dir}(\phi_z^B|\alpha_B\beta\beta^\top)] \tag{20}$$

$$= \sum_{z=1}^{K}\left(\ln\Gamma(\alpha_B) - \sum_{k,k'\in[1,K]}\ln\Gamma(\alpha_B\beta_k\beta_{k'}) + \sum_{k,k'\in[1,K]}(\alpha_B\beta_k\beta_{k'}-1)E[\ln\phi^B_{z,(k,k')}]\right). \tag{21}$$

Here $\sum_{i\in[1,K]}\beta_i = 1$, and then $L_{\mathrm{rules}-K}(\beta_{1:K-1},\beta_K)$ are useful to differentiate $L_{\mathrm{rules}-K}(\beta)$ such as

$$\frac{\partial L_{\mathrm{rules}}}{\partial \beta_k} = \frac{\partial L_{\mathrm{rules}-K}}{\partial \beta_k} + \frac{\partial L_{\mathrm{rules}-K}}{\partial \beta_K}\frac{\partial \beta_K}{\partial \beta_k}, \tag{22}$$

$$= \frac{\partial L_{\mathrm{rules}-K}}{\partial \beta_k} - \frac{\partial L_{\mathrm{rules}-K}}{\partial \beta_K}, \tag{23}$$

$$= \sum_{z=1}^{K}\left[-2\alpha_B\sum_{k'\in[1,K]}\beta_{k'}\Psi(\alpha_B\beta_k\beta_{k'}) + \sum_{k'\in[1,K]}\alpha_B\beta_{k'}\Big\{E[\ln\phi^B_{z,(k,k')}] + E[\ln\phi^B_{z,(k',k)}]\Big\}\right]$$
$$- \sum_{z=1}^{K}\left[-2\alpha_B\sum_{k'\in[1,K]}\beta_{k'}\Psi(\alpha_B\beta_K\beta_{k'}) + \sum_{k'\in[1,K]}\alpha_B\beta_{k'}\Big\{E[\ln\phi^B_{z,(K,k')}] + E[\ln\phi^B_{z,(k',K)}]\Big\}\right], \tag{24}$$

$$= \alpha_B\sum_{z=1}^{K}\sum_{k'\in[1,K]}\beta_{k'}\left[2\left\{\Psi(\alpha_B\beta_k\beta k') - \Psi(\alpha_B\beta_K\beta k')\right\} + E\Big[\ln\frac{\phi^B_{z,(k,k')}\phi^B_{z,(k',k)}}{\phi^B_{z,(K,k')}\phi^B_{z,(k',K)}}\Big]\right]. \tag{25}$$

Digamma function $\Psi(\gamma)$ meets that

$$\Psi(\gamma) = \ln(\gamma) - \frac{1}{2\gamma} + \sum_{n=1}^{\infty}\frac{\zeta(1-2n)}{\gamma^{2n}},\ \zeta(x) = \sum_{n=1}^{\infty}\frac{1}{n^x}, \tag{26}$$

where ζ is a Riemann's zeta function. By application of an approximation $\Psi(\gamma) \simeq \ln \gamma$ to (25),

$$\frac{\partial L_{\text{rules}}}{\partial \beta_k} = \alpha_B \sum_{z=1}^{K} \sum_{k' \in [1,K]} \beta_{k'} \left[2 \ln(\beta_k / \beta_K) + \ln E\Big[\frac{\phi^B_{z,(k,k')} \phi^B_{z,(k',k)}}{\phi^B_{z,(K,k')} \phi^B_{z,(k',K)}} \Big] \right], \tag{27}$$

because

$$E[\ln x_i] = \Psi(\alpha_i) - \Psi\Big(\sum_{\forall j} \alpha_j \Big) \simeq \ln \alpha_i - \ln \Big(\sum_{\forall j} \alpha_j \Big) = \ln E[x_i], \tag{28}$$

such that $x \sim \mathrm{Dir}(x; \alpha)$.

Summary

In summary,

$$\frac{\partial L}{\partial \beta_k} = -\frac{\alpha - 1}{T_K} - \sum_{i=k+1}^{K-1} \frac{1}{T_i} + \alpha_B \sum_{z=1}^{K} \sum_{k' \in [1,K]} \beta_{k'} \left[2 \left\{ \Psi(\alpha_B \beta_k \beta k') - \Psi(\alpha_B \beta_K \beta k') \right\} + E\Big[\ln \frac{\phi^B_{z,(k,k')} \phi^B_{z,(k',k)}}{\phi^B_{z,(K,k')} \phi^B_{z,(k',K)}} \Big] \right], \tag{29}$$

for $k = 1, \cdots K-1$, and $\partial L / \partial \beta_K = 0$.

References

[Liang2009] Percy Liang, Michael I. Jordan, Dan Klein, "Probabilistic grammars and hierarchical Dirichlet processes," The Oxford Handbook of Applied Bayesian Analysis, 2009.