Stochastic Natural Gradient Algorithm and Online learning for Latent Dirichlet Allocation

Nanbo Sun

Electrical & Computer Engineering, National University of Singapore, Singapore.
nanbosun@u.nus.edu

Abstract. This report derives in detail a stochastic optimization algorithm — stochastic natural gradient algorithm [1], and apply it to Latent Dirichlet Allocation. We first introduce the concept of stochastic gradient descent and natural gradient. Next, we combine them to create the stochastic natural gradient algorithm by using the information geometry. Finally, we apply the algorithm to Latent Dirichlet Allocation.

1 Stochastic Natural gradient algorithm

Let us introduce these concept by using an example of multilayer perceptron. Each **sample** z is a pair composed of an **input variable** x and a **desired output** y. The output of the network is a function $f(x, w)$ of input variable x and the **weight vector** w. And, there are L samples $\{z_i\}, i = 1 \dots L$.

1.1 Stochastic gradient descent

The performance of the network is measured by a **loss function** $Q(z, w)$. Usually, the loss function is the squared mean error which is defined as following:

$$Q_{mse}(z, w) \triangleq \frac{1}{2}(y - f(x, w))^2 \tag{1}$$

Then, the expected risks and empirical risk can be written as:

$$C(w) \triangleq \boldsymbol{E}_z Q(z, w) \triangleq \int Q(z, w)p(z)dz \approx \frac{1}{L}\sum_i Q(z_i, w) \tag{2}$$

where $p(z)$ is the *ground truth* distribution of sample z. In the case of the empirical risk, the expectation then reduces to a simple average $1/L \sum_i Q(z_i, w)$.

If we use *gradient descent* algorithm to optimize it, the update equation is:

$$w_{t+1} = w_t - \gamma_t \nabla_w C(w) \approx w_t - \gamma_t \sum_i \bigtriangledown_w Q(z_i, w_t) \tag{3}$$

If we use *stochastic gradient descent* algorithm to optimize it, the update equation is:

$$w_{t+1} = w_t - \gamma_t \nabla_w Q(z_t, w_t) \tag{4}$$

The algorithm presents the concept of *stochastic approximation*. It's hoped that Eq.4 behaves like Eq.3.

Only if it satisfies the following 4 conditions [2], we can prove that the stochastic gradient descent is convergent.

$$\forall \epsilon > 0, \inf_{(w-w^*)^2 > \epsilon} (w - w^*) \nabla_w C(w) > 0 \tag{5}$$

The hypothesis simply states that the opposite of the gradient always points toward the minimum w^*.

In order to obtain convergence, the parameter updates $\gamma_t \nabla_w Q(z, w)$ must become smaller and smaller when the parameter vector approaches the optimum w^*. The size of the gradients can be described with a first order expansion of $\nabla_w Q$ in w^*.

$$\boldsymbol{E}_z(\nabla_w Q(z,w)^2) \approx \boldsymbol{E}_z(\nabla_w Q(z,w^*)^2) + \boldsymbol{E}((w-w^*)\dots) < A + B(w-w^*)^2 \tag{6}$$

where $A, B \geq 0$ are constants. When $w \to w^*$, $\boldsymbol{E}_z(\nabla_w Q(z,w)^2) \to A$. Therefore, we must use *decreasing learning rates*

$$\sum \gamma_t^2 < \infty \tag{7}$$

Suppose the learning rates decrease so fast that $\sum \gamma_t = R < \infty$. This would effectively maintain the parameters within a certain radius of their initial value. Therefore

$$\sum \gamma_t = \infty \tag{8}$$

1.2 Natural Gradient

Assume $S = \{\boldsymbol{w} \in R^n\}$ is a parameter space and $C(w)$ is a function defined in the space. Then, the distance between $\boldsymbol{w}$ and $\boldsymbol{w} + d\boldsymbol{w}$ is defined as following:

$$\begin{aligned} \text{Euclidean space:} \quad & \|d\boldsymbol{w}\|^2 = \sum_{i=1}^{n} (dw_i)^2 \\ \text{Riemannian space:} \quad & \|d\boldsymbol{w}\|^2 = \sum_{i,j} g_{ij}(\boldsymbol{w}) dw_i dw_j \end{aligned}$$

where $G = (g_{ij})$ is the Riemannian metric tensor.

Theorem 1. *The steepest descent direction of $C(w)$ in Riemannian space is [3]*

$$-\tilde{\nabla} C(w) = -G^{-1}(w) \nabla C(w) \tag{9}$$

Proof. Assume $d\boldsymbol{w} = \epsilon \boldsymbol{a}$,we want to minimize $C(\boldsymbol{w} + d\boldsymbol{w})$.

$$\begin{aligned} &\min_{\boldsymbol{a}} C(\boldsymbol{w} + d\boldsymbol{w}) = C(\boldsymbol{w}) + \epsilon \nabla C(\boldsymbol{w}) \boldsymbol{a} \\ &\text{s.t.} \|\boldsymbol{a}\|^2 = \sum g_{ij} a_i a_j = 1 \end{aligned}$$

We can use the Lagrangian method to solve this problem

$$\frac{\partial}{\partial a_i}\{\nabla C(\boldsymbol{w})^T \boldsymbol{a} - \lambda \boldsymbol{a}^T G \boldsymbol{a}\} = 0$$
$$\boldsymbol{a} = \frac{1}{2\lambda} G^{-1} \nabla C(\boldsymbol{w})$$
$$\Rightarrow -\tilde{\nabla} C(w) = -G^{-1}(w) \nabla C(w)$$

1.3 Information Geometry

Assume $y = f(x, w) + \epsilon$, where ϵ is the Gaussian noise.

$$p(z, w) = C \exp[-\frac{(y - f(x, w))^2}{2\sigma^2}] \tag{10}$$

Because we hope that $p(z, w) \approx p(z)$ which is the *ground truth* distribution of z, we can minimize the following equation

$$\min_w D_{kl}(p(z) \| p(z, w)) = \boldsymbol{E}_z \log \frac{p(z)}{p(z, w)} = \frac{1}{\sigma^2} \boldsymbol{E}_z Q_{mse}(z, w) + \text{Constant} \tag{11}$$

According to the above equation, we can find that it is the same as minimizing the *mean square error* loss function. We can measure the distance between $\boldsymbol{w}$ and $\boldsymbol{w} + d\boldsymbol{w}$ by following equation

$$D_{kl}(\boldsymbol{w} \| \boldsymbol{w} + d\boldsymbol{w}) = D(\boldsymbol{w} \| \boldsymbol{w} + d\boldsymbol{w})|_{\boldsymbol{w}=\boldsymbol{w}+d\boldsymbol{w}} \tag{12}$$
$$+ d\boldsymbol{w} \cdot \frac{d}{d\boldsymbol{w}} D(\boldsymbol{w} \| \boldsymbol{w} + d\boldsymbol{w})|_{\boldsymbol{w}=\boldsymbol{w}+d\boldsymbol{w}} \tag{13}$$
$$+ \frac{1}{2} d\boldsymbol{w}^T [\frac{d^2}{d\boldsymbol{w_i} d\boldsymbol{w_j}} D(\boldsymbol{w} \| \boldsymbol{w} + d\boldsymbol{w})]|_{\boldsymbol{w}=\boldsymbol{w}+d\boldsymbol{w}} d\boldsymbol{w} \tag{14}$$
$$= \frac{1}{2} d\boldsymbol{w}^T \underbrace{[\frac{d^2}{d\boldsymbol{w_i} d\boldsymbol{w_j}} D(\boldsymbol{w} \| \boldsymbol{w} + d\boldsymbol{w})]}_{\text{Fisher Information}} |_{\boldsymbol{w}=\boldsymbol{w}+d\boldsymbol{w}} d\boldsymbol{w} \tag{15}$$

It is obvious that the Fisher Information can be regarded as a Rimannian metric tensor G.

1.4 Stochastic Natural Gradient Algorithm

Instead of using the gradient, we can use the natural gradient to substitute it. Therefore, we can get the following equation

$$w_{t+1} = w_t - \gamma_t G^{-1}(w_t) \nabla_w Q(z_t, w_t) \tag{16}$$

where $G = \int -(\nabla_w^2 \log p(z, w)) p(z, w) dw$.

2 Online Learning for LDA [4]

Now, let's apply the stochastic gradient descent algorithm to Latent Dirichlet Allocation. From Eq.(11) in [4], we know that

$$\mathcal{L}(g, \boldsymbol{\lambda}) \triangleq D\boldsymbol{E}_g[l(n, \gamma(n, \boldsymbol{\lambda}), \phi(n, \boldsymbol{\lambda}), \boldsymbol{\lambda})|\lambda]. \tag{17}$$

Using the stochastic natural gradient descent, we get

$$\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} + \rho_t D G^{-1} \nabla_\lambda l(n_t, \gamma_t, \phi_t, \lambda) \tag{18}$$

According to Eq.(14), we know that $G^{-1}\nabla_\lambda l(n_t, \gamma_t, \phi_t, \lambda) = -\lambda_{kw}/D + \eta/D + n_{tm}\phi_{twk}$. Then, the update equation becomes

$$\begin{aligned}
\boldsymbol{\lambda} &\leftarrow \boldsymbol{\lambda} + \rho_t D(-\lambda_{kw}/D + \eta/D + n_{tm}\phi_{twk}) && (19)\\
&\leftarrow \boldsymbol{\lambda} + \rho_t(-\lambda_{kw} + \underbrace{\eta + D n_{tm}\phi_{twk}}_{\tilde{\lambda}_{kw}}) && (20)\\
&\leftarrow (1-\rho_t)\boldsymbol{\lambda} + \rho_t \tilde{\boldsymbol{\lambda}} && (21)
\end{aligned}$$

The above equation is the same as that in **Algorithm 2**.

References

1. Bottou, Lon, and Noboru Murata. "Stochastic approximations and efficient learning." The Handbook of Brain Theory and Neural Networks, Second edition,. The MIT Press, Cambridge, MA (2002).
2. Bottou, Lon. "Online learning and stochastic approximations." On-line learning in neural networks 17.9 (1998): 25.
3. Amari, Shun-Ichi. "Natural gradient works efficiently in learning." Neural computation 10.2 (1998): 251-276.
4. Hoffman, Matthew, Francis R. Bach, and David M. Blei. "Online learning for latent dirichlet allocation." advances in neural information processing systems. 2010.