Natural Gradient Method

Raghav K Singhal

NYU Courant

2018

Recap - Gradient Descent

Gradient Descent tries to minimize a function $f(\theta)$ by taking steps of size α in the direction of the steepest descent,

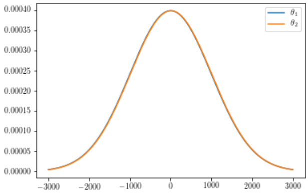
$$\theta_{n+1} = \theta_n - \alpha \nabla_\theta f(\theta_n)$$

where $\nabla_\theta f(\theta_n)$ is the direction of steepest descent:

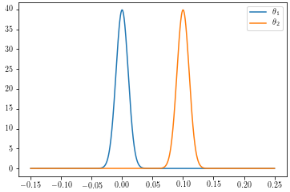
$$\lim_{\epsilon \to 0} \operatorname*{arg\,min}_{d\theta : ||d\theta||_2^2 \leq \epsilon^2} f(\theta + d\theta) = \nabla_\theta f(\theta)$$

Problems with Euclidean Distance

Let $p_\theta = \mathcal{N}(x|\mu, \sigma^2)$, where $\theta = [\mu, \sigma^2]$.
Now for $\theta_1 = [0, 1000]^T$ and $\theta_2 = [10, 1000]^T$, the euclidean distance between θ_1, θ_2 is 10, but they completely overlap.



However, for $\theta_1 = [0, 0.01]^T$ and $\theta_2 = [0.1, 0.01]^T$, they barely overlap but the euclidean distance between them is 0.1,



Information Geometry

The idea here is that in general the Euclidean metric is not the appropriate metric or distance function in parameter space.

We could use KL-divergence, but its not symmetric. However if we take a "local" view and analyze the geometry imposed by KL-Divergence, we realize that its FAKE NEWS!

KL Divergence

A Really "CLOSE" look.

We can show that KL-Divergence is 'locally' symmetric. Let p_θ be any smooth parametrized distribution. Now, we define the Fisher information matrix $G(\theta)$ as follows:

$$\begin{aligned} G(\theta) &= \mathbb{E}_{p_\theta}[\nabla_\theta \log p_\theta(x) \nabla_\theta \log p_\theta(x)^T] \\ &= -\mathbb{E}_{p_\theta}[\nabla_\theta^2 \log p_\theta] \end{aligned}$$

Then note that by taking a Taylor Series expansian around θ

$$\begin{aligned} \mathrm{KL}(p_{\theta+d\theta}(x)||p_\theta(x)) &= \frac{1}{2} d\theta^T G(\theta) d\theta + \mathcal{O}(||d\theta||^3) \\ \mathrm{KL}(p_\theta(x)||p_{\theta+d\theta}(x)) &= \frac{1}{2} d\theta^T G(\theta) d\theta + \mathcal{O}(||d\theta||^3) \end{aligned}$$

Therefore, $\mathrm{KL}(p_{\theta+d\theta}(x)||p_\theta(x)) = \mathrm{KL}(p_\theta(x)||p_{\theta+d\theta}(x))$. Hence, KL-Divergence is locally symmetric.

Natural Gradients

Now, the directioin of steepest descent in this geometry, induced by the Fisher metric, is as follows:

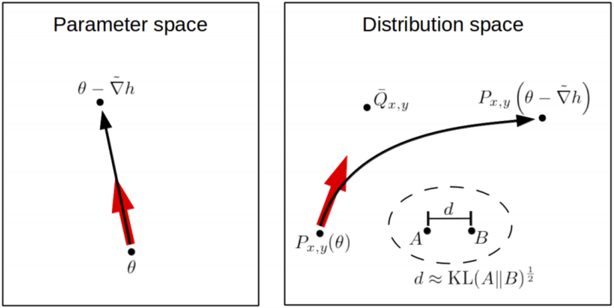
$$\lim_{\epsilon \to 0} \operatorname*{arg\,min}_{d\theta : KL(\theta || \theta + d\theta) \leq \epsilon} f(\theta + d\theta) = G(\theta)^{-1} \nabla_\theta f(\theta)$$

where we define $G(\theta)^{-1} \nabla_\theta f(\theta)$ as the natural gradient $\tilde{\nabla}_\theta f(\theta)$.

So gradient descent in this space is as follows:

$$\begin{aligned} \theta_{n+1} &= \theta_n - \alpha \tilde{\nabla}_\theta f(\theta_n) \\ &= \theta_n - \alpha G(\theta)^{-1} \nabla_\theta f(\theta_n) \end{aligned}$$

Information Geometry



The red arrow is the natural gradient direction, given by the vector $G(\theta)^{-1}\nabla h(\theta)$ in parameter space and the black arrow is the path generated by taking $\theta - \alpha G(\theta)^{-1}\nabla h$.

The Fisher

This formulation induces a Riemannian Geometry and we can view the Fisher matrix as inducing a norm in the distribution space:

$$||p_\theta||_{G_\theta} = \langle p_\theta, G(\theta)p_\theta \rangle$$

so this naturally gives us notions of length, geodesics, etc.

A geodesic is the shortest path between two points, for example in a Euclidean Space, the shortest distance between two points is a straight line, however in Riemannian Manifolds, the shortest distance between two points can be curved and is not unique.

Properties of Natural Gradients

1. We have now formulated the gradient descent algorithm in the space of prediction functions or distributions rather than parameters
2. When the natural gradient descent algorithm approches the optimum, i.e. $\mathbb{P}_\theta$ approches $\mathbb{Q}$, the Fisher matrix $G(\theta)$ approaches the true hessian of the loss function, $\mathbb{E}_{\mathbb{Q}}[\nabla^2 f(\theta)]$.
3. However, getting the actual Fisher Matrix is infeasible in most settings. So some authors suggest using the following approximation:

$$\tilde{G}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta p_\theta(x_i) \nabla_\theta p_\theta(x_i)^T$$

Note that this approximation is different from our previous formulation, as $\tilde{G}(\theta) = \mathbb{E}_{\hat{\mathbb{Q}}}[\nabla_\theta p_\theta(x) \nabla_\theta p_\theta(x)^T]$.