Information Geometry Optimisation Algorithms

Phan Trung Hai Nguyen

School of Computer Science
University of Birmingham
Birmingham B15 2TT
http://www.cs.bham.ac.uk/∼pxn683/

October 09, 2017

Introduction

- A very general optimisation method, defined for arbitrary search space
- Taking the advantage of natural gradient on Riemannian spaces
- IGO flow viewed as an ordinary differential equation
- Different IGO algorithms derived by approximating the IGO flow in different ways

Preliminary

- Search space $\mathcal{X}$ (e.g. binary search space $\{0,1\}^n$)
- P_θ family of probability distribution on $\mathcal{X}$
- Parameter $\theta = (\theta_1, \ldots, \theta_K) \in \Theta$, where Θ is space of parameter
- Given a function $L : \Theta \to \mathbb{R}$
- Problem

$$\arg\max_{\theta \in \Theta} L(\theta)$$

Gradient ascent

- Basic calculus:

$$\nabla L(\theta) = 0.$$

- If analytic solution not exist, resort to numerical methods.
- Gradient ascent: update current $\theta^{(t)}$ to a new value $\theta^{(t+\delta t)}$ such that

$$L(\theta^{(t)}) \leq L(\theta^{(t+\delta t)}).$$

- Update scheme:

$$\theta^{(t+\delta t)} \leftarrow \theta^{(t)} + \gamma \, \nabla_\theta L(\theta)\big|_{\theta=\theta^{(t)}},$$

where γ is learning rate, and vector of gradient

$$\nabla L(\theta) = \left(\frac{\partial L}{\partial \theta_1}, \frac{\partial L}{\partial \theta_2}, \ldots, \frac{\partial L}{\partial \theta_K} \right)$$

evaluated at $\theta = \theta^{(t)}$.

Disadvantages

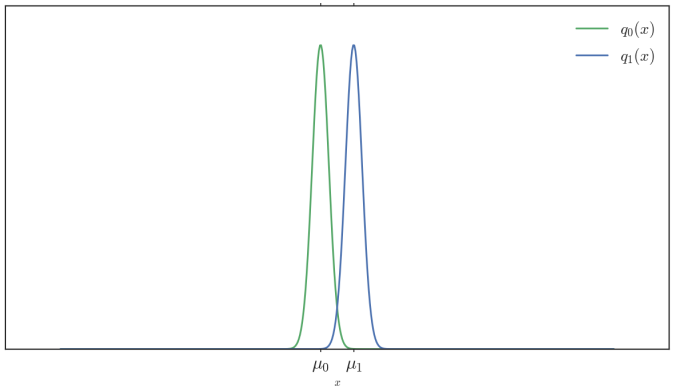
- Changes in parameters and function values measured by the Euclidean distance.
- Distance between $(\theta_1, \ldots, \theta_K)$ and $(\theta_1^*, \ldots, \theta_K^*)$ is given by

$$\sqrt{(\theta_1 - \theta_1^*)^2 + \ldots + (\theta_K - \theta_K^*)^2}$$

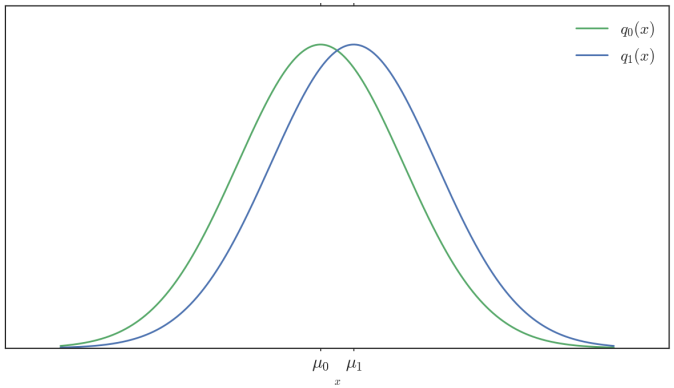
- However, most spaces are not Euclidean, but Riemannian.
- If Euclidean distance between parameters does not reflect the corresponding variation in the function values, then the (vanilla) gradient does not represent the steepest direction.

An example: Gaussian distributions

- Consider two following pairs of Gaussian distributions



$\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_1^2)$



$\mathcal{N}(\mu_1, \sigma_2^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$

- In both distributions, Euclidean distance between parameters

$$\sqrt{(\mu_1 - \mu_2)^2 + (\sigma_1^2 - \sigma_1^2)^2} = |\mu_1 - \mu_2|.$$

- The left two distributions less overlap compared to the right two distributions.
- Comparing parameters ignores the structure of the probability distributions, not naturally living in Euclidean but Riemannian space.

Fisher Information

- Measure distance between two probability distributions: Kullback-Leibler Divergence (KL, Kullback & Leibler (1951))

$$\mathrm{KL}(P_{\theta_1}||P_{\theta_2}) = \int_x \ln \frac{P_{\theta_1}(x)}{P_{\theta_2}(x)} P_{\theta_1}(x) dx = \mathbb{E}_{P_{\theta_1}} \left[\ln \frac{P_{\theta_1}(x)}{P_{\theta_2}(x)} \right]$$

- In case θ_1 and θ_2 are close, then

$$\mathrm{KL}(P_{\theta_1}||P_{\theta_2}) \approx \frac{1}{2}(\theta_1 - \theta_2)^T F_{\theta_2}(\theta_1 - \theta_2)$$

- F is the Fisher Information Matrix (Ollivier et al. 2017)

$$F_{\theta_2} = \nabla^2_{\theta_1} \mathrm{KL}(P_{\theta_1}||P_{\theta_2})\big|_{\theta_1=\theta_2}$$

- Fisher information matrix is a common Riemannian metric.

Natural gradient

- Using Fisher information matrix $F = (F_{ij})$, distance between $(\theta_1, \ldots, \theta_K)$ and $(\theta_1^*, \ldots, \theta_K^*)$ measured by

$$\sum_{1 \leq i,j \leq K} F_{ij}(\theta_i - \theta_i^*)(\theta_j - \theta_j^*)$$

- The vanilla gradient is replaced by natural gradient (Amari 1998)

$$\tilde{\nabla}_\theta L(\theta) = F^{-1}(\theta)\, \nabla_\theta L(\theta)$$

- New update rule for gradient ascent

$$\theta^{(t+\delta t)} \leftarrow \theta^{(t)} + \gamma\, \tilde{\nabla}_\theta L(\theta)\big|_{\theta=\theta^{(t)}}.$$

IGO flow

- Optimising a function $f : \mathcal{X} \to \mathbb{R}$
- Approach
  - Transforming f from search space $\mathcal{X}$ to parameter space Θ,
  - Defining $L(\theta)$ as the P_θ-average of the transformed function f',
  - Performing gradient ascent for $L(\theta)$ in space Θ,
- θ converges to a point s.t. samples from P_θ yield good values of f.
- Natural Evolution Strategy (NES, Wierstra et al. (2008)) considers

$$L(\theta) := \mathbb{E}_{P_\theta} f = \int_x f(x) P_\theta(x) dx.$$

- Unstable: expected value influenced by extreme values.

IGO flow (cont'd)

- Replacing f with a monotone rewriting $W^f_{\theta^t}$, depending on current parameter θ^t.
- Define $L(\theta) := \mathbb{E}_{P_\theta} W^f_{\theta^t} = \int_x P_\theta(x) W^f_{\theta^t}(x) dx$
- Continuous-time trajectory of θ

$$\begin{aligned}
\frac{d\theta^t}{dt} = \tilde{\nabla}_\theta L(\theta)\big|_{\theta=\theta^t} &= \tilde{\nabla}_\theta \int_x P_\theta(x) W^f_{\theta^t}(x) dx \Big|_{\theta=\theta^t} \\
&= \int_x W^f_{\theta^t}(x) \tilde{\nabla}_\theta P_\theta(x) dx \Big|_{\theta=\theta^t} \\
&= \int_x W^f_{\theta^t}(x) F^{-1}(\theta) \nabla_\theta (\ln P_\theta(x)) P_\theta(x) dx \Big|_{\theta=\theta^t} \\
&= F^{-1}(\theta^t) \int_x W^f_{\theta^t}(x) \frac{\partial \ln P_\theta(x)}{\partial \theta} \Big|_{\theta=\theta^t} P_{\theta^t}(x) dx,
\end{aligned}$$

where $\tilde{\nabla}_\theta P_\theta = P_\theta \tilde{\nabla}_\theta \ln P_\theta$ (log-likelihood trick)

IGO Algorithms

- Derived by approximating IGO flow in slightly different ways.
- Approximation often involves:
  - Discretising the IGO flow: $\theta^{t+\delta t} \approx \theta^t + \delta t \, \frac{d\theta^t}{dt}$ (δt: learning rate),
  - Sampling $(x_i)_{i=1}^N \sim P_{\theta^t}$, then calculating $\left(\nabla_\theta \ln P_\theta(x_i)\big|_{\theta=\theta^t}\right)$ and $\left(W_{\theta^t}^f(x_i)\right)$, then

$$G(\theta^t) := \int_x W_{\theta^t}^f(x) \nabla_\theta \ln P_\theta(x)\big|_{\theta=\theta^t} P_{\theta^t}(x) dx$$

$$\approx \frac{1}{N} \sum_{i=1}^N W_{\theta^t}^f(x_i) \nabla_\theta \ln P_\theta(x_i)\big|_{\theta=\theta^t}$$

  - Fisher information matrix, evaluated at $\theta = \theta^t$

$$F(\theta^t) \approx \frac{1}{N} \sum_{i=1}^N \left(\nabla_\theta \ln P_\theta(x_i)\right) \left(\nabla_\theta \ln P_\theta(x_i)\right)^T .$$

- The update scheme becomes

$$\theta^{(t+\delta t)} = \theta^{(t)} + \delta t \; F^{-1}(\theta^t) \; G(\theta^t)$$

References

Amari, S. (1998), 'Natural gradient works efficiently in learning', *Neural Computation* **10**(2), 251–276.
URL: *https://doi.org/10.1162/089976698300017746*

Kullback, S. & Leibler, R. A. (1951), 'On information and sufficiency', *Ann. Math. Statist.* **22**(1), 79–86.
URL: *https://doi.org/10.1214/aoms/1177729694*

Ollivier, Y., Arnold, L., Auger, A. & Hansen, N. (2017), 'Information-geometric optimization algorithms: A unifying picture via invariance principles', *J. Mach. Learn. Res.* **18**(1), 564–628.
URL: *http://dl.acm.org/citation.cfm?id=3122009.3122027*

Wierstra, D., Schaul, T., Peters, J. & Schmidhuber, J. (2008), Natural evolution strategies, *in* '2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)', pp. 3381–3387.