EM Algorithm for Entropy Search

A Preprint

John Doe[*]

Department of Computer Science
Cranberry-Lemon University
Pittsburgh, PA 15213
hippo@cs.cranberry-lemon.edu

August 4, 2022

Abstract

dah

Keywords First keyword · Second keyword · More

1 Idea

1.1 Entropy Search Lower Bound

The acquisition function for entropy search is defined as

$$\alpha_{\mathrm{ES}}(x) = \mathbb{H}[X^*|\mathcal{D}_t] - \mathbb{E}_{p(y|x)}\mathbb{H}[X^*|\mathcal{D}_t, y, x],$$

where x is the new sample point, X^* is a random variable of global minimizer x^* induced by the Gaussian process posterior, Y is a random variable representing the function value at x, $\mathcal{D}_t$ is our current data set. Alternatively, we can represent the acquisition function in a mutual information form,

$$\begin{aligned}\alpha_{\mathrm{ES}}(x) &= \mathbb{H}[X^*|\mathcal{D}_t] - \mathbb{H}[X^*|\mathcal{D}_t, Y, x]\\ &= \mathbb{H}[X^*] - \mathbb{H}[X^*|Y; x]\\ &= I[X^*, Y; x].\end{aligned}$$

We drop the $\mathcal{D}_n$ notation since it is fixed until the next sampling step. x is considered as a parameter for Y. (like a θ) Before achieving a lower bound for the acquisition function, let us introduce a lemma.

Lemma 1.1. *For given density function p and arbitrary density function q, we have*

$$\mathbb{E}_{p(x,y)}[log(p(x|y)] \geq \mathbb{E}_{p(x,y)}[log(q(x|y)].$$

The equality holds if and only if $p(x|y) = q(x|y)$.

Proof. By the concavity of log function and Jensen's inequality, we have $\mathrm{KL}(p(x|y)||q(x|y)) \geq 0$. The equality holds if and only if $p(x|y) = q(x|y)$. Break the fraction of the KL divergence and shift the negative term to the right,

$$\int p(x|y)\log(p(x|y))dx \geq \int p(x|y)\log(q(x|y))dx.$$

Take expectation on both sides,

$$\mathbb{E}_{p(y)}\int p(x|y)\log(p(x|y))dx \geq \mathbb{E}_{p(y)}\int p(x|y)\log(q(x|y))dx.$$

[*]footnote

Apply the Bayes formula,

$$\mathbb{E}_{p(x,y)}[\log(p(x|y)] \geq \mathbb{E}_{p(x,y)}[\log(q(x|y)].$$

□

A lower bound of the mutual information can be attained as following,

$$\begin{aligned}
\alpha_{\mathrm{ES}}(x) &= \mathbb{H}[X^*] - \mathbb{H}[X^*|Y;x] \\
&= \mathbb{H}[X^*] + \mathbb{E}_{p(y;x)}[p(x^*|y;x)\log(p(x^*|y;x))] \\
&= \mathbb{H}[X^*] + \mathbb{E}_{p(x^*,y;x)}[\log(p(x^*|y;x))] \\
&\geq \mathbb{H}[X^*] + \mathbb{E}_{p(x^*,y;x)}[\log(q(x^*|y;x))].
\end{aligned}$$

The last equality holds if and only if $q(x^*|y) = p(x^*|y)$ by Lemma 1.1. Since $\mathbb{H}[X^*]$ is independent with either q or x, we drop it and define the entropy search lower bound (ESLB) as

$$\mathrm{ESLB}(q,x) := \mathbb{E}_{p(x^*,y;x)}[\log(q(x^*|y))].$$

Because our goal is to find x that maximize the acquisition function $\alpha_{\mathrm{ES}}(x)$, we can instead maximize its lower bound ESLB with updating the variational posterior $q(x^*|y)$; see Algorithm 1.

Algorithm 1: EM Entropy Search (EM-ES)

Input: Sample set $\mathcal{D}_t$, initial x_0, kernel k
Output: acquisition maximizer x

1: initialize x_0
2: **for** n = 1:N **do**
3: E-step: update posterior $q_n(x^*|y) \leftarrow p(x^*|y;x_{n-1})$;
4: M-step: update parameter $x_n \leftarrow \arg\max_x \mathrm{ESLB}(q_n,x)$;
5: **end for**
6: return x_N

1.2 E-Step

For the expectation step, we update the posterior $p(x^*|y)$ such that the new posterior $q_n(x^*|y)$ is same with $p(x^*|y;x_{n-1})$. However, the distribution $p(x^*|y;x_{n-1})$ is still unknown and intractable to get a closed form. Instead, we apply Expected Improvement to construct a surrogate distribution $\tilde{q}(x^*|y)$ based on $\{\mathcal{D}_t \cup (x_{n-1},y)\}$. As a reminder, the expected improvement acquisition function is

$$\begin{aligned}
\alpha_{\mathrm{EI}}(x) &= \mathbb{E}_y[\max\{y_t^* - y, 0\}|\mathcal{D}_t] \\
&= (y_t^* - \mu(x))\Phi(\frac{y_t^* - \mu(x)}{\sigma(x)}) + \sigma(x)\phi(\frac{y_t^* - \mu(x)}{\sigma(x)}),
\end{aligned}$$

where y_t^* is the minimal value at current step t, $\mu(x)$ and $\sigma(x)$ are Gaussian process mean and standard deviation, $\phi(\cdot)$ and $\Phi(\cdot)$ are pdf/cdf of standard normal distribution. To distinguish, we let $\mu_t(x), \sigma_t(x)$ represent mean and variance of GP from $\mathcal{D}_t$, and $\mu_{t,n}(x), \sigma_{t,n}(x)$ denote GP from $\{\mathcal{D}_t \cup (x_n,y)\}$. Eventually, we have surrogate density function $\tilde{q}(x^*)$ as

$$\tilde{q}(x^*) = \frac{1}{C_{n-1}}\big((y_t^* - \mu_{t,n-1}(x^*))\Phi(\frac{y_t^* - \mu_{t,n-1}(x^*)}{\sigma_{t,n-1}(x^*)}) + \sigma_{t,n-1}(x^*)\phi(\frac{y_t^* - \mu_{t,n-1}(x^*)}{\sigma_{t,n-1}(x^*)})\big),$$

where C is a constant for normalization. We assign the new approximate posterior $q_n(x^*|y) = \tilde{q}(x^*)$.

1.3 M-Step

We have an alternative expression for ESLB in the M-step,

$$\begin{aligned}
\mathrm{ESLB}(q_n,x) &= \mathbb{E}_{p(x^*,y;x)}[\log(q_n(x^*|y))] \\
&= \mathbb{E}_{p(y;x)}[\mathbb{E}_{p(x^*|y)}[\log(q_n(x^*|y))]] \\
&= \mathbb{E}_{p(y;x)}[\mathbb{E}_{p(x^*|y)}[\log(\tilde{q}(x^*))]].
\end{aligned}$$

We can apply stochastic gradient descent by letting

$$p(x^*|y) = \frac{1}{C}\big((y_t^* - \mu_{t,x,y}(x^*))\Phi(\frac{y_t^* - \mu_{t,x,y}(x^*)}{\sigma_{t,x,y}(x^*)}) + \sigma_{t,x,y}(x^*)\phi(\frac{y_t^* - \mu_{t,x,y}(x^*)}{\sigma_{t,x,y}(x^*)})\big),$$

where $\mu_{t,x,y}$ and $\sigma_{t,x,y}$ are GP mean and std of $\mathcal{D}_t \cup (x, y)$. The constants C can be ignored for both $p(x^*|y)$ and $\tilde{q}(x^*|y)$.

References