Supplement to: That was fast! Speeding up NN search of high dimensional distributions.

Emanuele Coviello ECOVIELL@UCSD.EDU
Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093, USA

Adeel Mumtaz ADEELMUMTAZ@GMAIL.COM
Department of Computer Science
City University of Hong Kong
Kowloon Tong, Hong Kong

Antoni B. Chan ABCHAN@CITYU.EDU.HK
Department of Computer Science
City University of Hong Kong
Kowloon Tong, Hong Kong

Gert R.G. Lanckriet GERT@ECE.UCSD.EDU
Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093, USA

This is supplemental material for the manuscript "That was fast! Speeding up NN search of high dimensional distributions" (Coviello et al., 2013), appearing on the *Proceedings of the 30th International Conference on Machine Learning*, Atlanta, Georgia, USA, 2013. (JMLR: W&CP volume 28).

It consists of a collection of notes that fills in interesting details omitted in (Coviello et al., 2013) due to space limitations.

In Section 1 we fill in the derivation of the lower bounds. In Section 2 we illustrate how ℓ_m is not always monotonic and prove its convexity (which is used in Lemma 1 of (Coviello et al., 2013) to prove the monotonicity of ℓ_m^+). In Section 3 we elaborate on the approximation made in our branch and bound algorithm, and on how it affects the decisions of our algorithm.

1. Derivation of the lower bounds.

The lower bound presented in Section 4.2.2 of (Coviello et al., 2013), i.e., ℓ_q and ℓ_m, are easily derived from Hershey and Olsen (2007) variational approximation to the KL divergence between mixture models:

$$D(\mathcal{A}||\mathcal{B}) \approx \sum_i \pi_i \log \frac{\sum_{i'} \pi_{i'} \exp\{-D(\mathcal{A}_i||\mathcal{A}_{i'})\}}{\sum_j \omega_j \exp\{-D(\mathcal{A}_i||\mathcal{B}_j)\}}. \tag{1}$$

where $\mathcal{A} = \{\pi_i, \mathcal{A}_i\}$ is a mixture with weights π_i and components $\mathcal{A}_i$, and $\mathcal{B} = \{\omega_j, \mathcal{B}_j\}$ is a mixture with weights ω_j and components $\mathcal{B}_j$.

The lower bound to $D(\Theta||\mathcal{Q})$ is found for $\mathcal{A} = \Theta = \{\theta, \mathcal{M}, (1-\theta), \mathcal{Q}\}$ and $\mathcal{B} = \{1, \mathcal{Q}\}$:

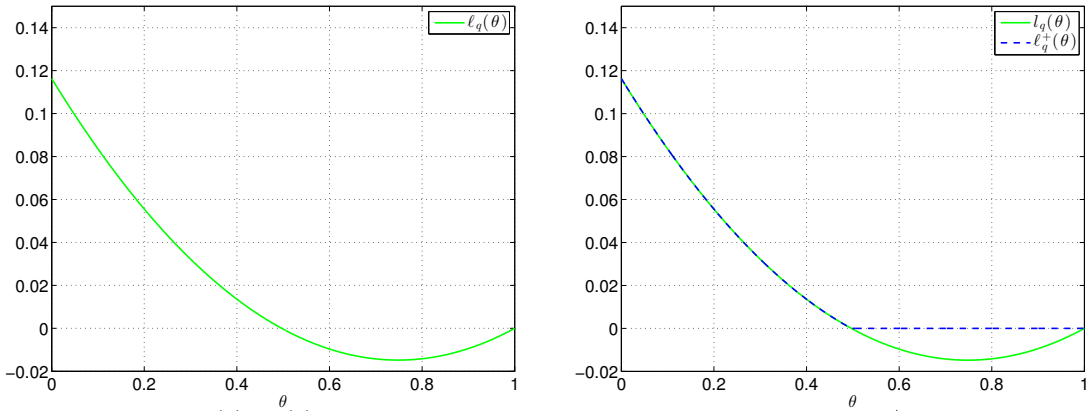
$$\begin{aligned}
\ell_q &= \theta \log \frac{\theta \exp\{-D(\mathcal{M}||\mathcal{M}) + (1-\theta)\exp\{-D(\mathcal{M}||\mathcal{Q})\}}{\exp\{-D(\mathcal{M}||\mathcal{Q})\}} \\
&\quad + (1-\theta)\log \frac{\theta\exp\{-D(\mathcal{Q}||\mathcal{M})\} + (1-\theta)\exp\{-D(\mathcal{Q}||\mathcal{Q})}{\exp\{-D(\mathcal{Q}||\mathcal{Q})\}} \\
&= \theta \log \frac{\theta + (1-\theta)\exp\{-D(\mathcal{M}||\mathcal{Q})\}}{\exp\{-D(\mathcal{M}||\mathcal{Q})\}} + (1-\theta)\log\left[\theta\exp\{-D(\mathcal{Q}||\mathcal{M})\} + 1 - \theta\right] \quad (2) \\
&= \theta \log\left[\theta\exp\{D(\mathcal{M}||\mathcal{Q})\} + (1-\theta)\right] + (1-\theta)\log\left[\theta\exp\{-D(\mathcal{Q}||\mathcal{M})\} + 1 - \theta\right] \quad (3)
\end{aligned}$$

where in (2) we use the fact that $D(\mathcal{Q}||\mathcal{Q}) = 0$, and in (3) we multiply numerator and denominator of the fist term inside the logarithm by $\exp\{D(\mathcal{M}||\mathcal{Q})\}$. Similarly, the lower bound to $D(\Theta||\mathcal{M})$ is found for $\mathcal{A} = \Theta = \{\theta, \mathcal{M}, (1-\theta), \mathcal{Q}\}$ and $\mathcal{B} = \{1, \mathcal{M}\}$:

$$\begin{aligned}
\ell_m &= \theta \log \frac{\theta \exp\{-D(\mathcal{M}||\mathcal{M}) + (1-\theta)\exp\{-D(\mathcal{M}||\mathcal{Q})\}}{\exp\{-D(\mathcal{M}||\mathcal{M})\}} \\
&\quad + (1-\theta)\log \frac{\theta\exp\{-D(\mathcal{Q}||\mathcal{M})\} + (1-\theta)\exp\{-D(\mathcal{Q}||\mathcal{Q})\}}{\exp\{-D(\mathcal{M}||\mathcal{Q})\}} \\
&= \theta \log\left[\theta + (1-\theta)\exp\{-D(\mathcal{M}||\mathcal{Q})\}\right] + (1-\theta)\log \frac{\theta\exp\{-D(\mathcal{Q}||\mathcal{M})\} + 1 - \theta}{\exp\{-D(\mathcal{Q}||\mathcal{M})\}} \quad (4) \\
&= \theta \log\left[\theta + (1-\theta)\exp\{-D(\mathcal{M}||\mathcal{Q})\}\right] + (1-\theta)\log\left[\theta + (1-\theta)\exp\{D(\mathcal{Q}||\mathcal{M})\}\right] \quad (5)
\end{aligned}$$

Note that, (3) and (5) hold as lower bounds, as explained in Section 4.2.2 of Coviello et al. (2013).

2. Monotonicity of lower bound



(a) $\ell_m(\theta)$ (b) $\ell_m(\theta)$ vs. $\ell_m^+(\theta)$

Figure 1: Add some caption

The function $\ell_m(\theta)$ is not always monotonic, as illustrated by the following example. Consider two histograms $\mathcal{Q}$ and $\mathcal{M}$ with bins:

$$q = \quad [0.2,\ 0.2,\ 0.15,\ 0.25,\ 0.2] \tag{6}$$

$$\mu = \quad [0.2,\ 0.2,\ 0.3,\ 0.1,\ 0.2] \tag{7}$$

for which we have $D(\mathcal{Q}||\mathcal{M}) \approx 0.1163$ and $D(\mathcal{M}||\mathcal{Q}) \approx 0.1251$.

In Figure 1(a), we plot $\ell_m(\theta)$ for $\theta \in [0\,1]$, which is visibly non monotonic. On the opposite, $\ell_m^+(\theta) = \max(0, \ell_m(\theta))$ is monotonic, as illustrated in Figure 1(b).

In general, Lemma 1 in (Coviello et al., 2013) gives a proof of the monotonicity of $\ell_m^+(\theta)$. Lemma 1 uses the convexity of $\ell_m(\theta)$, which follows from the positivity of the second derivative (for $\theta \in [0\,1]$). In particular, using $A = \exp\{-D(M||Q)\}$ and $B = \exp\{D(Q||M)\}$ to reduce clutter, we have:

$$\begin{aligned}
\frac{d}{d\theta}\ell_m(\theta) &= \frac{d}{d\theta}\theta\log[\theta + (1-\theta)A] + \frac{d}{d\theta}(1-\theta)\log[\theta + (1-\theta)B] \\
&= \log[\theta(1-A) + A] + \frac{\theta(1-A)}{\theta(1-A)+A} \\
&\quad + \frac{1-B}{\theta(1-B)+B} - \log[\theta(1-B)+B] + \frac{\theta(1-B)}{\theta(1-B)+B}
\end{aligned} \tag{8}$$

$$\begin{aligned}
\frac{d^2}{d\theta^2}\ell_m(\theta) &= \frac{d}{d\theta}\frac{d}{d\theta}\ell_m(\theta) \\
&= \frac{(1-A)[\theta(1-A)+2A]}{[\theta(1-A)+A]^2} + \frac{-(1-B)(1-B)}{[\theta(1-B)+B]^2} - \frac{(1-B)[\theta(1-B)+2B]}{[\theta(1-B)+B]^2} \\
&= \frac{(1-A)[\theta(1-A)+2A]}{[\theta(1-A)+A]^2} - \frac{(1-B)[(\theta+1)+B(1-\theta)]}{[\theta(1-B)+B]^2}.
\end{aligned} \tag{9}$$

Since $B = \exp\{D(Q||M)\} > 1$, $A = \exp\{-D(M||Q)\} \in [0\,1]$, for $\theta \in [0\,1]$, (9) is positive (from which follows the convexity).

3. Decisions of the approximated algorithm

In the derivation of Algorithm 2 in (Coviello et al., 2013), we use in sequence an approximation and a lower bound:

$$d_f(x_\theta, q) \approx \quad D(\Theta||\mathcal{Q}) \geq \quad \ell_q^+(\theta) \tag{10}$$

$$d_f(x_\theta, \mu) \approx \quad D(\Theta||\mathcal{M}) \geq \quad \ell_m^+(\theta) \tag{11}$$

In general, when (10) and (11) do not hold as equalities, the algorithm is subject to making two types of incorrect decisions, i.e., exploring parts of the search space that could instead be safely pruned (over-explorative behavior), or pruning away parts that should be explored (under-explorative behavior). Interestingly, we can argue that, instead of making one type of mistake or the other *randomly*, our algorithm is over-explorative on nodes to which the query is relatively close and under-explorative on nodes to which the query is further away (see (Coviello et al., 2013)).

Conjecture: If $q' = \mu' + \delta'$, for δ' small, we have $d_f(x_\theta, \mu) \geq l_m^+(\theta)$.

Proof: We have $x'_\theta = \mu' + (1-\theta)\delta'$ and $\mu' = x'_\theta - (1-\theta)\delta'$. Consider the Riemannian manifold around x'_θ (with curvature $\nabla f^*(x'_\theta)$), and that the Riemannian metrics associated to f and f^* have identical infinitesimal length (Amari, 2009; Nielsen and Nock, 2009). Consequently we have:[1]

$$d_f(x_\theta, \mu) = \frac{1}{2}(1-\theta)^2 \delta'^t \nabla^2 f^*(x'_\theta)\delta' \tag{12}$$

$$= \frac{1}{2}(1-\theta)^2 \delta^t \nabla^2 f(x_\theta)\delta = (1-\theta)^2\Delta \tag{13}$$

where we use the notation $\Delta = \frac{1}{2}\delta^t \nabla^2 f(x_\theta)\delta$ to reduce clutter. Similarly, we have that $d_f(q, \mu) = \frac{1}{2}\Delta$ and $d_f(\mu, q) = \frac{1}{2}\Delta$. Using the approximations $\exp\{a\} = 1+a$ and $\log a = a-1$ (for a small), we have that:

$$\ell_m(\theta) = (1\text{–}\theta)\log\left[1+(1\text{–}\theta)\Delta\right] + \theta\log\left[1\text{–}(1\text{–}\theta)\Delta\right] \tag{14}$$

$$\leq \log\left\{(1\text{–}\theta)\left[1+(1\text{–}\theta)\Delta\right] + \theta\left[1\text{–}(1\text{–}\theta)\Delta\right]\right\} \tag{15}$$

$$= \log\left\{1 + \left[(1-\theta)^2 - \theta(1-\theta)\right]\Delta\right\} \tag{16}$$

$$= \left[(1-\theta)^2 - \theta(1-\theta)\right]\Delta \leq (1-\theta)^2\Delta \tag{17}$$

where (15) follow from Jensen inequality and (17) form the fact that $\theta(1-\theta) \geq 0$ for $\theta \in [0\,1]$. Since $d_f(x_\theta, \mu) \geq 0$ we also have $d_f(x_\theta, \mu) \geq \ell_m^+(\theta)$.

Next, we illustrate the *over-explorative* behavior of our Algorithm 2 from (Coviello et al., 2013) when the approximations hold as lower bounds, i.e.:

$$d_f(x_\theta, q) \geq \ell_q^+(\theta), \tag{18}$$

$$d_f(x_\theta, \mu) \geq \ell_m^+(\theta), \tag{19}$$

$$\mathcal{L}(\theta) \geq \ell_{\mathcal{L}}(\theta) \equiv \ell_q^+(\theta) + \frac{\theta}{1-\theta}\left(\ell_m^+(\theta) - R\right) \tag{20}$$

In Step 4 of our algorithm when comparing $\ell_{\mathcal{L}}(\theta)$ to $c = d_f(x_c, q)$ (where x_c is the candidate NN), we can have the following situations:

- If $\mathcal{L}(\theta) \geq \ell_{\mathcal{L}} > c$ we *safely* prune the node;
- If $\mathcal{L}(\theta) > c > \ell_{\mathcal{L}}$ [loose lower bound] we *incorrectly* decide not to prune the node yet: we may end up exploring more and waste computation;
- If $c > \mathcal{L}(\theta) \geq \ell_{\mathcal{L}}$ [tight lower bound] we correctly not prune the node (yet).

For the rest of the comparisons (i.e., Steps 5, 6 and 7 of our algorithm), we can have:

- If $\ell_m < d_f(x_\theta, \mu) < R$ [tight lower bound]: we correctly assume x_θ is in the ball;
  - $\ell_q \leq d_f(x_\theta, q) \leq c$ [tight lower bound]: we *correctly* decide to explore (e.g., x_θ is better than candidate $\mathcal{X}_c$);

1. To show (12) we can use Legendre duality $d_f(x_\theta, \mu) = f(x_\theta) + f^*(\mu') - \langle \mu', x_\theta \rangle$ and second order Taylor expansion of $f^*(\mu') = f^*(x'_\theta - (1-\theta)\delta')$ around x'_θ.

  - $\ell_q \leq c \leq d_f(x_\theta, q)$ [loose lower bound]: we *incorrectly* decide to explore — even if we might have ended up exploring anyway based on later iterations, in general we may waste computations;
  - $c \leq \ell_q \leq d_f(x_\theta, q)$: we *safely* update θ_l;

- If $\ell_m < R < d_f(x_\theta, \mu)$ [loose lower bound] we *incorrectly* assume x_θ is in the ball
  - $\ell_q \leq d_f(x_\theta, q) \leq c$ [tight lower bound]: we *incorrectly* decide to explore: wasteful;
  - $\ell_q \leq c \leq d_f(x_\theta, q)$ [loose lower bound]: we *incorrectly* decide to explore: wasteful;
  - $c \leq \ell_q \leq d_f(x_\theta, q)$: we *incorrectly* update θ_r instead of θ_l: this dilate the effective size of the ball and may incur in wasteful explorations;

- If $R < \ell_m < d_f(x_\theta, \mu)$ we *safely* assume x_θ is not the ball, and update θ_r.

On the opposite, when the query and the node are far away, the approximations can actually hold as *upper* bounds, which has the opposite effect of making the the algorithm under-explorative. Since the left-sided centroid $x_\theta = \arg\max_c \theta d_f(x, \mu) + (1-\theta) d_f(x, q)$ is zero forcing (Nielsen and Nock, 2009), it will have smaller support than the mixture model with modes μ and q, and consequently $d_f(x_\theta, \mu)$ (respectively, $d_f(x_\theta, q)$) will be smaller than $D(\Theta||\mathcal{M})$ (respectively, $D(\Theta||\mathcal{Q})$). If ℓ_q^+ and ℓ_m^+ are not tight, (10) and (11) will hold as upper bounds.

References

S. Amari. Information geometry and its applications: convex function and dually flat manifold. *Emerging Trends in Visual Computing*, pages 75–102, 2009.

E. Coviello, A. Mumtaz, A.B. Chan, and G.R.G. Lanckriet. That was fast! Speeding up NN search of high dimensional distributions. In *Proceedings of the 25th international conference on Machine learning*, 2013.

J.R. Hershey and P.A. Olsen. Approximating the Kullback Leibler divergence between Gaussian mixture models. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4. Ieee, 2007. ISBN 1424407273.

F. Nielsen and R. Nock. Sided and symmetrized Bregman centroids. *IEEE Transactions on IT*, 2009.