Boosted Density Estimation Remastered

Zac Cranko[1,2] and Richard Nock[2,1,3]

[1]The Australian National University [2]CSIRO Data61 [3]The University of Sydney

Quick Summary

- Learn a density function incrementally
- Use *classifiers* for the incremental updates (similar to GAN discriminators)
- Unlike other state of the art attempts, achieve strong convergence results (geometric) using a weak learning assumption on the classifiers (in the paper!)

$$\sup_{D:\mathcal{X}\to(0,1)} \mathrm{E}_{Q_0}[\log D] - \mathrm{E}_P[\log(1-D)]$$

Take $f(t) \stackrel{\text{def}}{=} t \log t - (t+1)\log(t+1)$ and $\varphi(D) \stackrel{\text{def}}{=} \frac{D}{1-D}$. Then

$$\begin{aligned}
&\sup_{D:\mathcal{X}\to(0,1)} \mathrm{E}_{Q_0}[\log D] - \mathrm{E}_P[\log(1-D)] \\
&\qquad = \sup_{D:\mathcal{X}\to(0,1)} \mathrm{E}_{Q_0}[f' \circ \varphi \circ D] - \mathrm{E}_P[f^* \circ f' \circ \varphi \circ D] \\
&\qquad = \sup_{d:\mathcal{X}\to(0,\infty)} \mathrm{E}_{Q_0}[f' \circ d] - \mathrm{E}_P[f^* \circ f' \circ d] \\
&\qquad = \mathrm{E}_{Q_0}\left[f' \circ \frac{\mathrm{d}P}{\mathrm{d}Q_0}\right] - \mathrm{E}_P\left[f^* \circ f' \circ \frac{\mathrm{d}P}{\mathrm{d}Q_0}\right]
\end{aligned}$$

Take $f(t) \stackrel{\text{def}}{=} t \log t - (t+1)\log(t+1)$ and $\varphi(D) \stackrel{\text{def}}{=} \frac{D}{1-D}$. Then

$$\begin{aligned}
\sup_{D:\mathcal{X}\to(0,1)} & \mathrm{E}_{Q_0}[\log D] - \mathrm{E}_P[\log(1-D)] \\
&= \sup_{D:\mathcal{X}\to(0,1)} \mathrm{E}_{Q_0}[f' \circ \varphi \circ D] - \mathrm{E}_P[f^* \circ f' \circ \varphi \circ D] \\
&= \sup_{d:\mathcal{X}\to(0,\infty)} \mathrm{E}_{Q_0}[f' \circ d] - \mathrm{E}_P[f^* \circ f' \circ d] \\
&= \mathrm{E}_{Q_0}\left[f' \circ \frac{\mathrm{d}P}{\mathrm{d}Q_0}\right] - \mathrm{E}_P\left[f^* \circ f' \circ \frac{\mathrm{d}P}{\mathrm{d}Q_0}\right]
\end{aligned}$$

Recall:

$$\forall f: \int f(x) P(\mathrm{d}x) = \int f(x) \frac{\mathrm{d}P}{\mathrm{d}Q_0}(x) Q_0(\mathrm{d}x)$$

Main Idea

$$d_1 \in \underset{d':\mathcal{X}\to(0,\infty)}{\arg\max}\ \mathrm{E}_{Q_0}[f' \circ d'] - \mathrm{E}_P[f^* \circ f' \circ d']$$

1. Find d_1 as above

Main Idea

$$d_1 \in \underset{d':\mathcal{X}\to(0,\infty)}{\arg\max}\ \mathrm{E}_{Q_0}[f' \circ d'] - \mathrm{E}_P[f^* \circ f' \circ d']$$

1. Find d_1 as above
2. Multiply $d_1(x)Q_0(\mathrm{d}x)$ to find $P(\mathrm{d}x)$

Main Idea

$$d_1 \in \underset{d':\mathcal{X}\to(0,\infty)}{\arg\max}\ \mathrm{E}_{Q_0}[f' \circ d'] - \mathrm{E}_P[f^* \circ f' \circ d']$$

1. Find d_1 as above
2. Multiply $d_1(x)Q_0(\mathrm{d}x)$ to find $P(\mathrm{d}x)$
3. Finished. Get a job at a hedge fund next door

Unfortunately this is not so simple since in practice we can only approximately solve the maximisation.

Main Idea

$$d_1 \in \underset{d':\mathcal{X}\to(0,\infty)}{\arg\max}\ \mathrm{E}_{Q_0}[f' \circ d'] - \mathrm{E}_P[f^* \circ f' \circ d']$$

1. Find d_1 as above
2. Multiply $d_1(x)Q_0(\mathrm{d}x)$ to find $P(\mathrm{d}x)$
3. Finished. Get a job at a hedge fund next door

Unfortunately this is not so simple since in practice we can only approximately solve the maximisation. Sadface.

Solution

$$d_t \in \operatorname*{arg\,max}_{d':\mathcal{X}\to(0,\infty)} \mathrm{E}_{Q_{t-1}}[f' \circ d'] - \mathrm{E}_P[f^* \circ f' \circ d']$$

$$\tilde{Q}_t(\mathrm{d}x) = d_t^{\alpha_t}(x) \cdot \tilde{Q}_{t-1}(\mathrm{d}x), \quad Q_t = \frac{1}{Z_t}\tilde{Q}_t, \quad \text{where} \quad Z_t \stackrel{\text{def}}{=} \int \mathrm{d}\tilde{Q}_t,$$

1. Some step size parameters $\alpha_t \in (0, 1)$
2. Treat the updates as *classifiers* $d_t = \exp \circ c_t$

Solution

$$d_t \in \operatorname*{arg\,max}_{d':\mathcal{X}\to(0,\infty)} \mathrm{E}_{Q_{t-1}}[f' \circ d'] - \mathrm{E}_P[f^* \circ f' \circ d']$$

$$\tilde{Q}_t(\mathrm{d}x) = d_t^{\alpha_t}(x) \cdot \tilde{Q}_{t-1}(\mathrm{d}x), \quad Q_t = \frac{1}{Z_t}\tilde{Q}_t, \quad \text{where} \quad Z_t \stackrel{\mathrm{def}}{=} \int \mathrm{d}\tilde{Q}_t,$$

1. Some step size parameters $\alpha_t \in (0, 1)$
2. Treat the updates as *classifiers* $d_t = \exp \circ c_t$
   - The classifiers are distinguishing between samples originating from P and Q_{t-1} like in a GAN
   - However unlike a GAN there is not necessarily a simple fast sampler for Q_{t-1}, but there is a closed-form density function

Solution

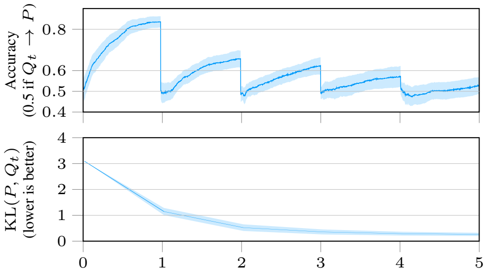
$$d_t \in \underset{d':\mathcal{X}\to(0,\infty)}{\arg\max}\ \mathrm{E}_{Q_{t-1}}[f' \circ d'] - \mathrm{E}_P[f^* \circ f' \circ d']$$

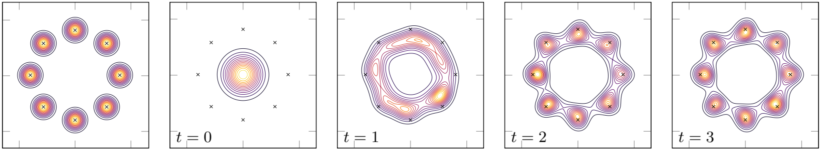
$$\tilde{Q}_t(\mathrm{d}x) = d_t^{\alpha_t}(x) \cdot \tilde{Q}_{t-1}(\mathrm{d}x), \quad Q_t = \frac{1}{Z_t}\tilde{Q}_t, \quad \text{where} \quad Z_t \overset{\text{def}}{=} \int \mathrm{d}\tilde{Q}_t,$$

1. Some step size parameters $\alpha_t \in (0,1)$
2. Treat the updates as *classifiers* $d_t = \exp \circ c_t$
   - The classifiers are distinguishing between samples originating from P and Q_{t-1} like in a GAN
   - However unlike a GAN there is not necessarily a simple fast sampler for Q_{t-1}, but there is a closed-form density function

Convergence of $Q_t \to P$ in KL-divergence with a weak learning assumption on the updates as classifiers. With additional minimal assumptions: geometric convergence.

Experiments





Thanks for listening, come chat to us at poster #161. (Bring beer!)