GAN Crash course

Eirikur Agustsson

March 21, 2018

f-divergences

Given two distributions P and Q with density $p(x)$ and $q(x)$ over a domain $\mathcal{X}$, we define the *f-divergence*,

f-divergences

Given two distributions P and Q with density $p(x)$ and $q(x)$ over a domain $\mathcal{X}$, we define the *f-divergence*,

$$D_f(P\|Q) = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x, \tag{1}$$

where the function $f : \mathbb{R}_+ \to \mathbb{R}$ is a convex (lower-semicontinuous) function, with $f(1) = 0$.

f-divergences

Given two distributions P and Q with density $p(x)$ and $q(x)$ over a domain $\mathcal{X}$, we define the *f-divergence*,

$$D_f(P\|Q) = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x, \tag{1}$$

where the function $f : \mathbb{R}_+ \to \mathbb{R}$ is a convex (lower-semicontinuous) function, with $f(1) = 0$.
Example, set $f(u) = -\log(u) = \log(1/u)$, obtaining

$$D_f(P\|Q) = \int_{\mathcal{X}} q(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x = D_{\mathrm{KL}}(Q\|P), \tag{2}$$

which is the well known Kullback-Leibner divergence.

f-divergences

Other divergences that belong to the f divergence family:

Name	$D_f(P\|Q)$	Generator $f(u)$
Kullback-Leibler	$\int p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x$	$u \log u$
Reverse KL	$\int q(x) \log \frac{q(x)}{p(x)} \, \mathrm{d}x$	$-\log u$
Pearson χ^2	$\int \frac{(q(x)-p(x))^2}{p(x)} \, \mathrm{d}x$	$(u-1)^2$
Squared Hellinger	$\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \, \mathrm{d}x$	$\left(\sqrt{u}-1\right)^2$
Jensen-Shannon	$\frac{1}{2} \int p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)} \, \mathrm{d}x$	$-(u+1) \log \frac{1+u}{2} + u \log u$

f-divergences

Other divergences that belong to the f divergence family:

Name	$D_f(P\|Q)$	Generator $f(u)$
Kullback-Leibler	$\int p(x)\log\frac{p(x)}{q(x)}\,\mathrm{d}x$	$u\log u$
Reverse KL	$\int q(x)\log\frac{q(x)}{p(x)}\,\mathrm{d}x$	$-\log u$
Pearson χ^2	$\int \frac{(q(x)-p(x))^2}{p(x)}\,\mathrm{d}x$	$(u-1)^2$
Squared Hellinger	$\int \left(\sqrt{p(x)}-\sqrt{q(x)}\right)^2\,\mathrm{d}x$	$\left(\sqrt{u}-1\right)^2$
Jensen-Shannon	$\frac{1}{2}\int p(x)\log\frac{2p(x)}{p(x)+q(x)}+q(x)\log\frac{2q(x)}{p(x)+q(x)}\,\mathrm{d}x$	$-(u+1)\log\frac{1+u}{2}+u\log u$

Expressive family, but note that the form is constrained an integral of a pointwise function between p and q.

f-divergences

Other divergences that belong to the f divergence family:

Name	$D_f(P\|Q)$	Generator $f(u)$
Kullback-Leibler	$\int p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x$	$u \log u$
Reverse KL	$\int q(x) \log \frac{q(x)}{p(x)} \, \mathrm{d}x$	$-\log u$
Pearson χ^2	$\int \frac{(q(x)-p(x))^2}{p(x)} \, \mathrm{d}x$	$(u-1)^2$
Squared Hellinger	$\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mathrm{d}x$	$\left(\sqrt{u}-1\right)^2$
Jensen-Shannon	$\frac{1}{2}\int p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)} \, \mathrm{d}x$	$-(u+1)\log \frac{1+u}{2} + u \log u$

Expressive family, but note that the form is constrained an integral of a pointwise function between p and q.
Cannot capture e.g. Wasserstein distance:

$$W(P,Q) = \inf\{E_{(x,y)\sim R}[c(x,y)] \Big| x \sim P, y \sim Q\}, \tag{3}$$

which measure minimal expected cost $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ over couplings R between P and Q.

Standard GAN

It turns out the standard GAN can be related to the JS-divergence.

Standard GAN

It turns out the standard GAN can be related to the JS-divergence. Recall, that we then have a discriminator D and a generator G playing the following two-player minimax game:

$$\min_G \max_D \quad \mathbb{E}_{x \sim P_{\text{data}}}[\log D(x)] + \mathbb{E}_{x \sim G(z)}[\log(1 - D(x)))] \tag{4}$$

$$= \min_G \max_D \quad \int_x p_{\text{data}}(x) \log(D(x)) + p_{G(z)}(x) \log(1 - D(x)) dx \tag{5}$$

where z has a fixed prior distribution (e.g. uniform).

Standard GAN

It turns out the standard GAN can be related to the JS-divergence.
Recall, that we then have a discriminator D and a generator G playing the following two-player minimax game:

$$
\begin{aligned}
&\min_G \max_D \quad \mathbb{E}_{x \sim P_{\text{data}}}[\log D(x)] + \mathbb{E}_{x \sim G(z)}[\log(1 - D(x)))] && (4) \\
&= \min_G \max_D \quad \int_x p_{\text{data}}(x) \log(D(x)) + p_{G(z)}(x) \log(1 - D(x)) dx && (5)
\end{aligned}
$$

where z has a fixed prior distribution (e.g. uniform).
Now, since $a \log(y) + b \log(1 - y)$ is maximized at $y = a/(a + b)$, we know that the optimal D is $D^*_G(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_{G(z)}(x)}$.

Standard GAN

Now suppose when maxmizing D we find D^*_G. Then we can compute

$$
\begin{aligned}
&\min_G \max_D \mathbb{E}_{x\sim P_{\text{data}}}[\log D(x)] + \mathbb{E}_{x\sim G(z)}[\log(1-D(x))]] \\
&= \min_G \mathbb{E}_{x\sim P_{\text{data}}}[\log(\frac{p_{\text{data}}(x)}{p_{\text{data}}(x)+p_{G(z)}(x)})] + \mathbb{E}_{x\sim G(z)}[\log(\frac{p_{G(z)}(x)}{p_{\text{data}}(x)+p_{G(z)}(x)}))] \\
&= \min_G \int_x p_{\text{data}}(x)\log(\frac{p_{\text{data}}(x)}{\frac{1}{2}(p_{\text{data}}(x)+p_{G(z)}(x))}) \\
&\qquad + p_{G(z)}(x)\log(\frac{p_{G(z)}(x)}{\frac{1}{2}(p_{\text{data}}(x)+p_{G(z)}(x))})dx - \log(4) \\
&= \min_G D_{\mathrm{KL}}(p_{\text{data}}\|\tfrac{1}{2}(p_{\text{data}}+p_{G(z)})) + D_{\mathrm{KL}}(p_{\text{data}}\|\tfrac{1}{2}(p_{\text{data}}+p_{G(z)})) - \log(4) \\
&= \min_G 2D_{\text{Jenson-Shannon}}(p_{\text{data}}\|p_{G(z)}) - \log(4)
\end{aligned}
$$

So, in this case, G is minimizing the Jensen-Shannon divergence between p_{data} and $p_{G(z)}$!

Standard GAN

- So what is missing?

Standard GAN

- So what is missing?
- First, in practice, we only consider a limited family of functions

$$\min_{\theta \in \Theta} \max_{\omega \in \Omega} \quad \mathbb{E}_{x \sim P_{\text{data}}}[\log D_\omega(x)] + \mathbb{E}_{x \sim G_\theta(z)}[\log(1 - D_\omega(x)))], \tag{6}$$

where $\theta \in \Theta$ and $\omega \in \Omega$ parameterize families of functions (e.g. neural networks).

Standard GAN

- So what is missing?
- First, in practice, we only consider a limited family of functions

$$\min_{\theta \in \Theta} \max_{\omega \in \Omega} \quad \mathbb{E}_{x \sim P_{\text{data}}}[\log D_\omega(x)] + \mathbb{E}_{x \sim G_\theta(z)}[\log(1 - D_\omega(x)))], \tag{6}$$

  where $\theta \in \Theta$ and $\omega \in \Omega$ parameterize families of functions (e.g. neural networks).
- Since $D^*_G(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x)+p_{G(z)}(x)}$ depends on p_{data}, it is unlikely it is an element of the function family $\{D_\omega | \omega \in \Omega\}$.

Standard GAN

- So what is missing?
- First, in practice, we only consider a limited family of functions

$$\min_{\theta \in \Theta} \max_{\omega \in \Omega} \quad \mathbb{E}_{x \sim P_{\text{data}}}[\log D_\omega(x)] + \mathbb{E}_{x \sim G_\theta(z)}[\log(1 - D_\omega(x)))], \tag{6}$$

  where $\theta \in \Theta$ and $\omega \in \Omega$ parameterize families of functions (e.g. neural networks).
- Since $D_G^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_{G(z)}(x)}$ depends on p_{data}, it is unlikely it is an element of the function family $\{D_\omega | \omega \in \Omega\}$.
- Second, instead of finding the global mini-maximum, we only find a saddle point via double-loop SGD.
- Thus, we cannot simply plug D_G^* into the GAN objective!

Othe divergences

Can obtain any f-divergence (not just JS) by changing the loss

Othe divergences

Can obtain any f-divergence (not just JS) by changing the loss

$$\min_G \max_D \quad \mathbb{E}_{x\sim P_{\text{data}}}[\log D(x)] + \mathbb{E}_{x\sim G(z)}[\log(1 - D(x)))] \tag{7}$$

$$\tag{8}$$

Othe divergences

Can obtain any f-divergence (not just JS) by changing the loss

$$\min_G \max_D \quad \mathbb{E}_{x \sim P_{\text{data}}}[\log D(x)] + \mathbb{E}_{x \sim G(z)}[\log(1 - D(x)))] \tag{7}$$

$$\tag{8}$$

Becomes

$$\min_G \max_D \quad \mathbb{E}_{x \sim P_{\text{data}}}[f(D(x)] + \mathbb{E}_{x \sim G(z)}[g(D(x)] \tag{9}$$

$$\tag{10}$$

In particular, a notable alternative is the Least Squares GAN

Thank you for your attention!