f-divergence

$p(x), q(x)$分别表示x从分布P，Q中sample出来的概率。f可以是不同的函数，必须是凸函数，且$f(1) = 0$。f-divergence的式子如下，可以表示P和Q之间的差异，

$$D_f(P||Q) = \int_x q(x) f\left(\frac{p(x)}{q(x)}\right) dx$$

那么为什么这个式子可以表示P和Q之间的差异呢?

如果现在$q(x) = p(x)$，那么$D_f(P||Q) = 0$，表示q和p之间的距离为0.

如果q和p有一些很小的差距，算出来的divergence就大于0.

$$\int_x q(x) f\left(\frac{p(x)}{q(x)}\right) dx = 0 \geq f\left(\int_x q(x) \frac{p(x)}{q(x)} dx\right) = f(1) = 0$$

f-divergence P and Q are two distributions. $p(x)$ and q(x) are the probability of sampling x.

$$D_f(P||Q) = \int_x q(x) f\left(\frac{p(x)}{q(x)}\right) dx$$

f is convex

f(1) = 0

$D_f(P||Q)$ evaluates the difference of P and Q

If $p(x) = q(x)$ for all x

smallest $D_f(P||Q) = \int_x q(x) f\left(\frac{p(x)}{q(x)}\right) dx = 0$

($\frac{p(x)}{q(x)}$ = 1, $f(1)$ = 0)

$$D_f(P||Q) = \int_x q(x) f\left(\frac{p(x)}{q(x)}\right) dx$$

Because f is convex

$$\geq f\left(\int_x \cancel{q(x)} \frac{p(x)}{\cancel{q(x)}} dx\right)$$

$$= f(1) = 0$$

If P and Q are the same distributions, $D_f(P||Q)$ has the smallest value, which is 0

如果$f(x) = xlogx$，那么$D_f(P||Q)$就是KL divergence；

如果$f(x) = -logx$，那么$D_f(P||Q)$就是Reverse divergence；

如果$f(x) = (x-1)^2$，那么$D_f(P||Q)$就是Chi Square；

$f(x) = xlogx$

KL

$$D_f(P||Q) = \int_x q(x) \frac{p(x)}{q(x)} log\left(\frac{p(x)}{q(x)}\right) dx = \int_x p(x) log\left(\frac{p(x)}{q(x)}\right) dx$$

$f(x) = -logx$

Reverse KL

$$D_f(P||Q) = \int_x q(x) \left(-log\left(\frac{p(x)}{q(x)}\right)\right) dx = \int_x q(x) log\left(\frac{q(x)}{p(x)}\right) dx$$

$f(x) = (x-1)^2$

Chi Square

$$D_f(P||Q) = \int_x q(x) \left(\frac{p(x)}{q(x)} - 1\right)^2 dx = \int_x \frac{(p(x) - q(x))^2}{q(x)} dx$$

Fenchel Conjugate

每个凸函数都有一个Conjugate function f^*，是由x和$f(x)$导出来的。对于值的计算，我们可以通过穷举所有t的值代入，看到底哪个t可以使$xt - f(x)$的值最大。即

$$f^*(t) = \max_{x \in dom(f)} \{xt - f(x)\}$$

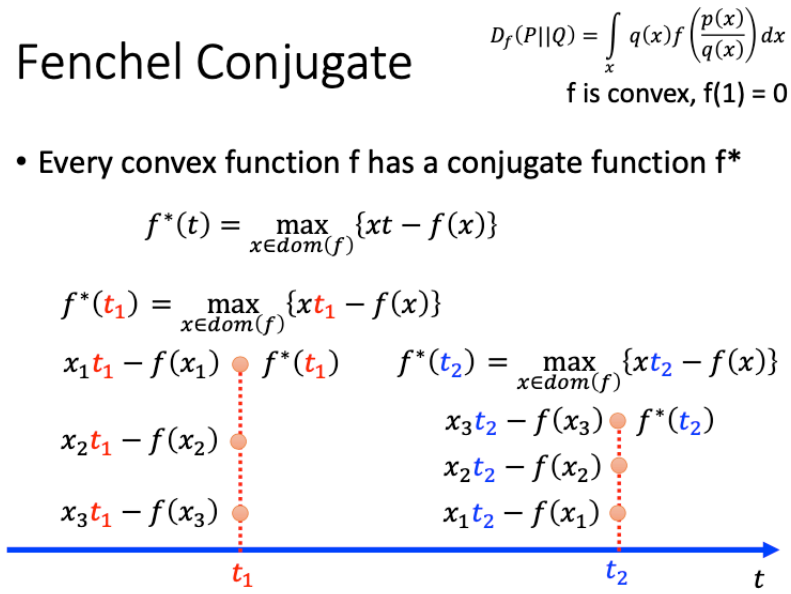
对于值的计算来举一个例子，当$t = t_1$时，

$$f^*(t) = \max_{x \in dom(f)} \{xt_1 - f(x)\}$$

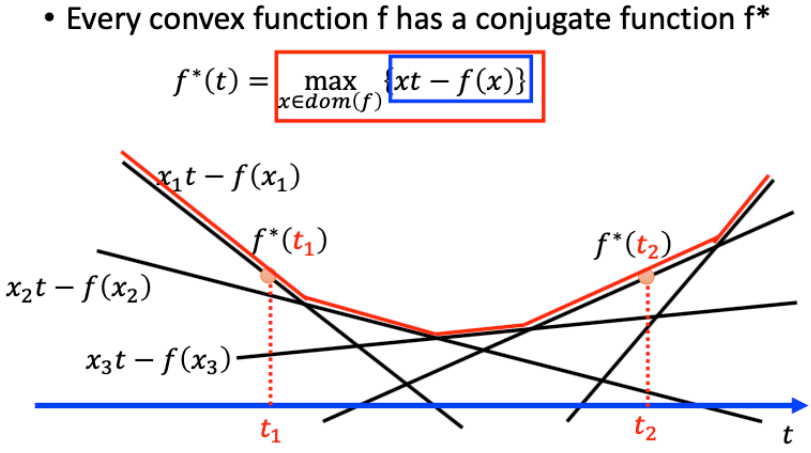
这时x的取值范围为$x = \{x_1, x_2, x_3\}$，代入x的值，计算$x_1t_1 - f(x_1), x_2t_1 - f(x_2), x_3t_1 - f(x_3)$的值，$f^*(t)$即为三者最大；

代入t_2的值来计算$f^(t_2)$的值；

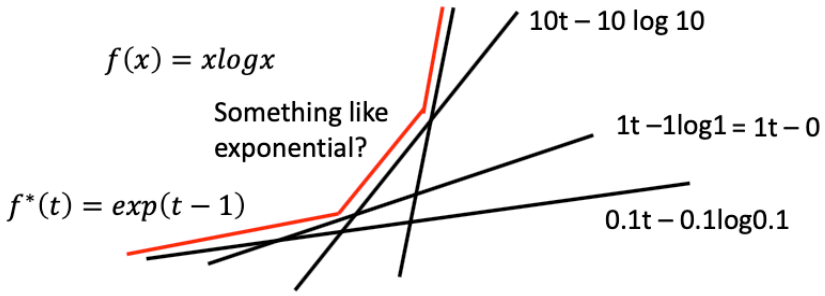
……



这样每个t值都带进去算很麻烦，因此出现了第二种方案，把$xt - f(x)$用图形描述出来。找出这些直线的upper bound，可以发现$f^*(t)$是凸函数。



现在假设$f(x) = xlogx$，代入$xt - f(x)$计算，画出$x = \{0.1, 1, 10, \dots\}$时的函数图像，并找出这些直线的upper bound，如下图所示，如果进行了很多次运算，这些直线的upper bound和$f^*(t) = exp(t-1)$的图像很接近。



下图是这个过程具体的证明，

- Every convex function f has a conjugate function f*
- (f*)* = f

$$f^*(t) = \max_{x \in dom(f)} \{xt - f(x)\}$$

$$f(x) = xlogx \longleftrightarrow f^*(t) = exp(t-1)$$

$$f^*(t) = \max_{x \in dom(f)} \{xt - xlogx\}$$

$g(x) = xt - xlogx$ Given t, find x maximizing $g(x)$

$t - logx - 1 = 0 \quad x = exp(t-1)$

$$f^*(t) = exp(t-1) \times t - exp(t-1) \times (t-1) = exp(t-1)$$

Connection with GAN

$$f^*(t) = \max_{x \in dom(f)} \{xt - f(x)\} \longleftrightarrow f(x) = \max_{t \in dom(f^*)} \{xt - f^*(t)\}$$

$$D_f(P||Q) = \int_x q(x) f\left(\frac{p(x)}{q(x)}\right) dx \qquad \frac{p(x)}{q(x)} \qquad \frac{p(x)}{q(x)}$$

$$= \int_x q(x) \left(\max_{t \in dom(f^*)} \left\{ \frac{p(x)}{q(x)} t - f^*(t) \right\} \right) dx$$

$$\approx \max_D \int_x p(x) D(x) dx - \int_x q(x) f^*(D(x)) dx$$

D is a function whose input is x, and output is t

$$D_f(P||Q) \geq \int_x q(x) \left(\frac{p(x)}{q(x)} D(x) - f^*(D(x)) \right) dx$$

$$= \int_x p(x) D(x) dx - \int_x q(x) f^*(D(x)) dx$$

$f(x)$与$f^*(x)$互为Conjugate function。把$x = \frac{p(x)}{q(x)}$代入，得

$$D_f(P||Q) = \int_x q(x) f\left\{ \max_{x \in dom(f)} \{ \frac{p(x)}{q(x)} t - f^*(t) \} \right\} dx$$

我们现在可以用一个discriminator D，来帮助我们求解这个max的问题，输入为x，输出就是满足条件的t，就不用穷举所有的t才能找到我们的最优解。如果用$D(x)$来替代x，就可以表示$D_f(P||Q)$的lower bound，即

$$D_f(P||Q) \geq \int_x q(x) f\left(\frac{p(x)}{q(x)} D(x) - f^*(D(x))\right) dx$$
$$= \int_x p(x) D(x) dx - \int_x q(x) f^*(D(x)) dx$$

如果随便找一个D，最后得出来的值肯定比divergence的值小；但如果是找一个最优（max）的D，预测出来的t就是最准的，就可以使结果逼近divergence，即

$$D_f(P||Q) \approx \max_D \int_x p(x) D(x) dx - \int_x q(x) f^*(D(x)) dx$$

$$D_f(P||Q) \approx \max_D \int_x p(x) D(x) dx - \int_x q(x) f^*(D(x)) dx$$
$$= \max_D \{E_{x\sim P}[D(x)] - E_{x\sim Q}[f^*(D(x))]\}$$

Samples from P　　Samples from Q

$$D_f(P_{data}||P_G) = \max_D \{E_{x\sim P_{data}}[D(x)] - E_{x\sim P_G}[f^*(D(x))]\}$$

$$G^* = arg \min_G D_f(P_{data}||P_G)$$

Original GAN has different V(G,D)

$$= arg \min_G \max_D \{E_{x\sim P_{data}}[D(x)] - E_{x\sim P_G}[f^*(D(x))]\}$$
$$= arg \min_G \max_D V(G,D)$$ familiar? ☺

$V(G,D)$可以有不同的形式，不同的divergence就有不同的$V(G,D)$。

下图中列出了不同的divergence和generator。

$$D_f(P_{data}||P_G) = \max_D\{E_{x\sim P_{data}}[D(x)] - E_{x\sim P_G}[f^*(D(x))]\}$$

Name	$D_f(P\|Q)$	Generator $f(u)$
Total variation	$\frac{1}{2}\int \lvert p(x)-q(x)\rvert\,\mathrm{d}x$	$\frac{1}{2}\lvert u-1\rvert$
Kullback-Leibler	$\int p(x)\log\frac{p(x)}{q(x)}\,\mathrm{d}x$	$u\log u$
Reverse Kullback-Leibler	$\int q(x)\log\frac{q(x)}{p(x)}\,\mathrm{d}x$	$-\log u$
Pearson χ^2	$\int \frac{(q(x)-p(x))^2}{p(x)}\,\mathrm{d}x$	$(u-1)^2$
Neyman χ^2	$\int \frac{(p(x)-q(x))^2}{q(x)}\,\mathrm{d}x$	$\frac{(1-u)^2}{u}$
Squared Hellinger	$\int \left(\sqrt{p(x)}-\sqrt{q(x)}\right)^2\,\mathrm{d}x$	$(\sqrt{u}-1)^2$
Jeffrey	$\int (p(x)-q(x))\log\left(\frac{p(x)}{q(x)}\right)\,\mathrm{d}x$	$(u-1)\log u$
Jensen-Shannon	$\frac{1}{2}\int p(x)\log\frac{2p(x)}{p(x)+q(x)} + q(x)\log\frac{2q(x)}{p(x)+q(x)}\,\mathrm{d}x$	$-(u+1)\log\frac{1+u}{2} + u\log u$
Jensen-Shannon-weighted	$\int p(x)\pi\log\frac{p(x)}{\pi p(x)+(1-\pi)q(x)} + (1-\pi)q(x)\log\frac{q(x)}{\pi p(x)+(1-\pi)q(x)}\,\mathrm{d}x$	$\pi u\log u - (1-\pi+\pi u)\log(1-\pi+\pi u)$
GAN	$\int p(x)\log\frac{2p(x)}{p(x)+q(x)} + q(x)\log\frac{2q(x)}{p(x)+q(x)}\,\mathrm{d}x - \log(4)$	$u\log u - (u+1)\log(u+1)$

Using the f-divergence you like ☺

https://arxiv.org/pdf/1606.00709.pdf

Name	Conjugate $f^*(t)$
Total variation	t
Kullback-Leibler (KL)	$\exp(t-1)$
Reverse KL	$-1-\log(-t)$
Pearson χ^2	$\frac{1}{4}t^2+t$
Neyman χ^2	$2-2\sqrt{1-t}$
Squared Hellinger	$\frac{t}{1-t}$
Jeffrey	$W(e^{1-t}) + \frac{1}{W(e^{1-t})} + t - 2$
Jensen-Shannon	$-\log(2-\exp(t))$
Jensen-Shannon-weighted	$(1-\pi)\log\frac{1-\pi}{1-\pi e^{t/\pi}}$
GAN	$-\log(1-\exp(t))$

可以用不同的divergence又有什么优点呢?

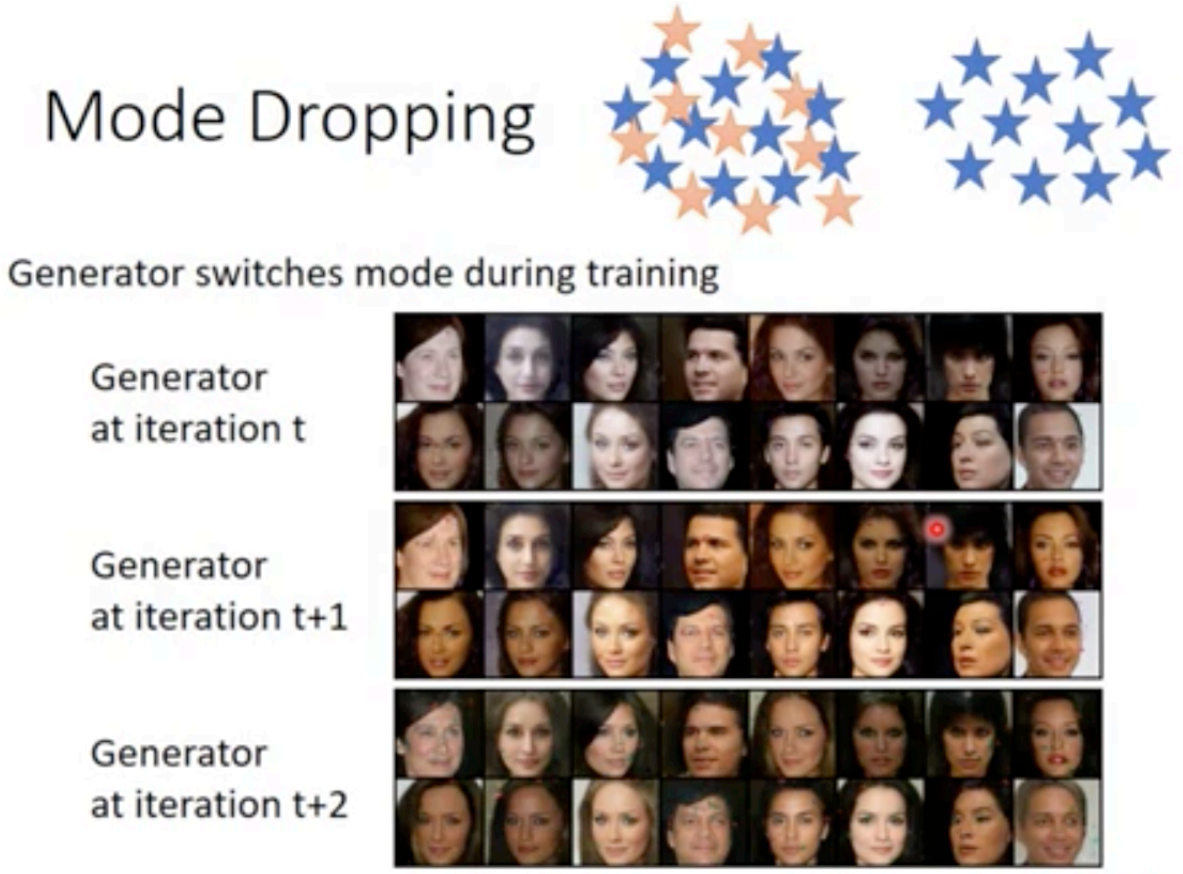
Mode Collapse

当我们在训练GAN的时候，可能会遇到mode collapse，real data的distribution是非常宽泛的，但generated data的distribution可能会非常小。比如我们在生成二次元人物的时候，可能会出现下图中的结果，某张特定的人脸开始蔓延，变得到处都是，同一张人脸会不断反复地出现。

Mode Dropping

mode dropping的情况比mode collapse要稍微简单一点，现在real data有两种不同的distribution，而generator只会产生一种distribution的数据。

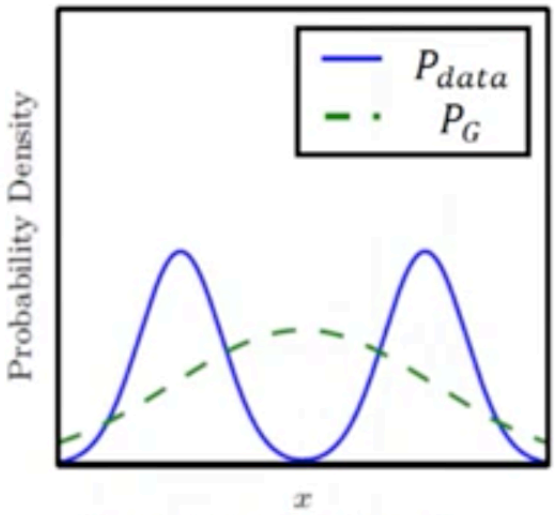
Generator第一次会先产生一些白皮肤的人，再进行一次generator，会产生一些黄皮肤的人，再进行一次generator，会产生一些黑皮肤的人。每次只产生一种分布的数据。



会出现这个问题，一个很可能的原因就是divergence选得不好。

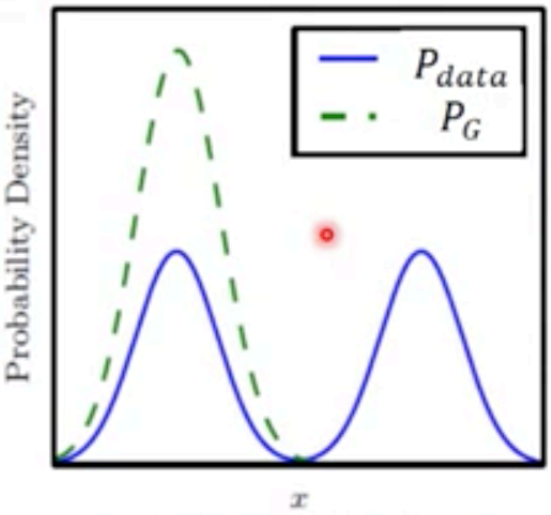
Flaw in Optimization?

$$KL = \int P_{data} log \frac{P_{data}}{P_G} dx$$



Maximum likelihood
(minimize KL($P_{data}||P_G$))

Reverse $KL = \int P_G log \frac{P_G}{P_{data}} dx$



Minimize KL($P_G||P_{data}$)
(reverse KL)