f-GAN
Обычный GAN минимизирует JS дивергенцию. Другие расстояния?

1. Wasserstein

2. $\mathrm{KL}(p^*(x) \| p_G(x))$, $\mathrm{KL}(p_G(x) \| p^*(x))$

3. Pearson χ^2: $\int \frac{(p_g(x) - p^*(x))^2}{p^*(x)} dx$

2 f-divergence:

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

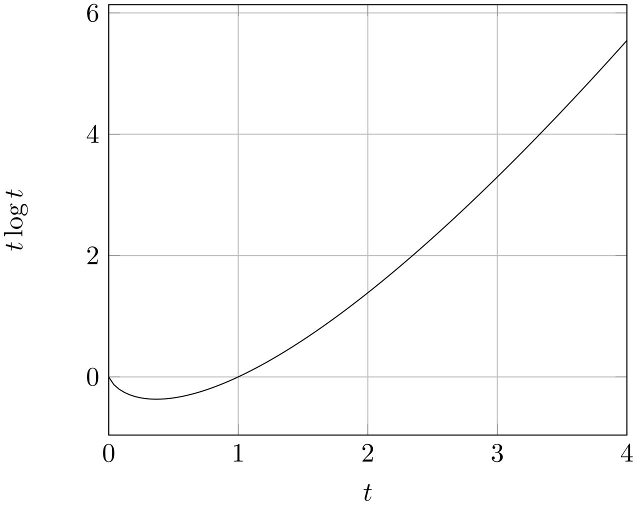
$$f : \mathbb{R}_+ \to \mathbb{R}, \text{выпукл.}, f(1) = 0$$

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

Не расстояние — нет неравенства треугольника.
Посмотрим, какие f соответствуют этим дивергенциям.

1. $\mathrm{KL}(q(x) \| p(x)) = D_f(p(x) \| q(x))$
$= \int q(x) \log \frac{q(x)}{p(x)} dx$, $f(t) = -\log t$

2. $\mathrm{KL}(p(x) \| q(x)) = D_f(p(x) \| q(x))$
$= \int p(x) \log \frac{p(x)}{q(x)} dx$, $f(t) = t \log t$



3.

$$JS(p(x)\|q(x)) = \frac{1}{2}\left(\int p(x)\log\left(\frac{p(x)}{\frac{p(x)+q(x)}{2}}\right) + \int q(x)\log\left(\frac{q(x)}{\frac{p(x)+q(x)}{2}}\right)\right) =$$

$$= \left[r(x) = \frac{p(x)}{q(x)}\right] =$$

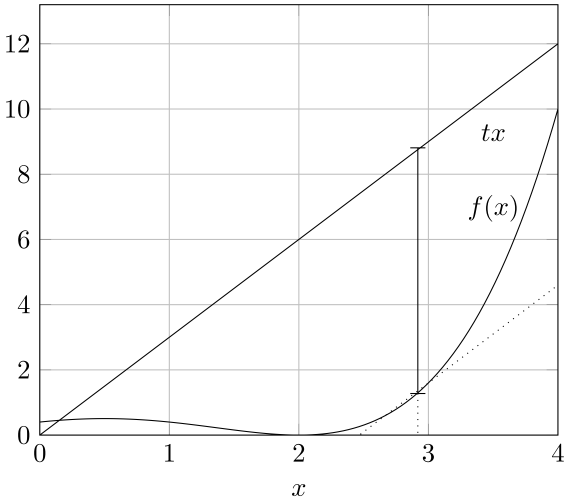
$$= \frac{1}{2}\left[\int q(x)\left[r(x)\left(\log\frac{r(x)}{r(x)+1} + \log r\right) + \log\left(\frac{1}{r(x)+1}\right) + \log 2\right]dx\right] =$$

$$f(t) = t\left(\log\frac{t}{t+1} + \log 2\right) + \log\frac{1}{t+1} + \log 2$$

Перейдем к обучению самих GAN'ов.

Рассмотрим сопряженные функции:

$$f^*(t) = \sup_{x\in\mathrm{dom} f}\{tx - f(x)\}$$



$$f^*(t) = \sup_{x\in\mathrm{dom} f}\{tx - f(x)\}$$

$$f(u) = \sup_{t}\{tu - f^*(t)\} = (f^*(t))^*(w)$$

$$
\begin{aligned}
D_f(P\|Q) &= \int q(x) f\left(\frac{p(x)}{q(x)}\right) dx = \\
&= \int q(x) \sup_t \left\{ t\frac{p(x)}{q(x)} - f^*(t) \right\} dx = \\
&= \int q(x) \left[t^*(x)\frac{p(x)}{q(x)} - f^*(t^*(x)) \right] dx = \\
&= \sup_{T(x)} \int q(x) \left[T(x)\frac{p(x)}{q(x)} - f(T(x)) \right] dx \geq \\
&\geq \sup_{T(x)\in\tau} \int (p(x)T(x) - q(x)f^*(T(x)))dx = \\
&= \sup_{T(x)\in\tau} \mathbb{E}_{p(x)}T(x) - \mathbb{E}_{q(x)}f^*(T(x))
\end{aligned}
$$

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

$$
p(x) = p^*(x), \quad q(x) = p_G(x), \quad T_w(x)
$$

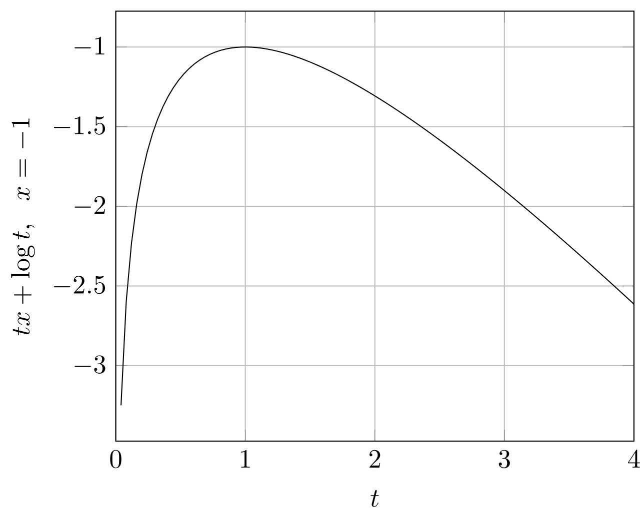
$$
\min_G \left[\max_w \left(\mathbb{E}_{p^*(x)} T_w(x) - \mathbb{E}_{p_G(x)} f^*(T_w(x)) \right) \right]
$$

Попробуем для прямого KL:

$$
\mathrm{KL}(p_G(x)\|p^*(x)) = D_f(p^*(x)\|p_G(x))
$$

$$
f(t) = -\log(t)
$$

$$
f^*(x) = \sup_t \{tx + \log t\}
$$

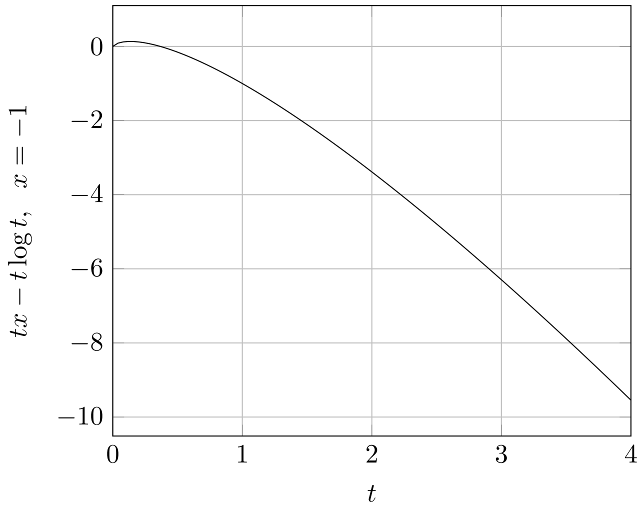


При $x < 0$ — $t = -\frac{1}{x}$, $f^*(x) = -(1 + \log(-x))$

$$L = \mathbb{E}_{p^*} T_w(x) + \mathbb{E}_{p_G}(1 + \log(-T_w(x)))$$
$$S_w(x); T_w(x) = -\exp(S_w(x))$$
$$L = -\mathbb{E}_{p^*} \exp(S_w(x)) + \mathbb{E}_{p_G} S_w(x)$$

Теперь для обратного KL:

$$f(t) = t \log t$$
$$f^*(x) = \sup_t tx - t \log t$$



$$t = \exp(x - 1)$$
$$f^* = \exp\{x - 1\}x - \exp\{x - 1\}(x - 1) = \exp\{x - 1\}$$
$$\mathbb{E}_{p^*} T_w(x) - \mathbb{E}_{p_{G(x)}} \exp(T_w(x) - 1)$$

Здесь фейковые семплы штрафуются намного сильнее.

Основные минусы такого анализа? Какие предположения были слишком сильными? Мы не оптимизируем $T_w(x)$ до конца; мы опираемся сильно на её оптимальность. Еще мы минимизируем нижнюю оценку по генератору. Это довольно плохо. Это приводит к очень нестабильному обучению и к многим проблемам, которые мы видим на практике.

Это может быть хороший мат. аппарат, но надо понимать, что он имеет свои ограничения.