Moreau-Yosida f-divergences

Dávid Terjék
dterjek@renyi.hu

Alfréd Rényi Institute of Mathematics
Budapest, Hungary

ICML 2021

Motivation

Kantorovich-Rubinstein formula:

$$W_1(\mu, \nu) = \sup_{f:X\to\mathbb{R}, \|f\|_L \le 1} \int f d\mu - \int f d\nu$$

Donsker-Varadhan formula:

$$D_{KL}(\mu\|\nu) = \sup_{f:X\to\mathbb{R}} \int f d\mu - \log \int e^f d\nu$$

What is this?

$$\sup_{f:X\to\mathbb{R}, \|f\|_L \le 1} \int f d\mu - \log \int e^f d\nu = ?$$

Moreau-Yosida approximation

(X,d) compact, $f : X \to \overline{\mathbb{R}}_+$ lower semicontinuous, $0 < \alpha, \lambda \in \mathbb{R}$:

$$f_{\lambda,\alpha}(x) = \min_{y \in X} f(y) + \lambda d(x,y)^\alpha$$

is the Moreau-Yosida approximation of index λ and order α of f.

Given a compact metric space (X,d), the metric space $(P(X), W_1)$ is compact, and $D_{KL}(\cdot \| \nu) : P(X) \to \overline{\mathbb{R}}_+$ is lower semicontinuous.

Kullback-Leibler divergence

$$D_{KL}(\mu \| \nu) = \begin{cases} \int \log \frac{d\mu}{d\nu} d\mu \text{ if } \mu \ll \nu, \\ \infty \text{ otherwise.} \end{cases}$$

Tight conjugate of Kullback-Leibler divergence with respect to ν:

$$D^*_{KL}(f \| \nu) = \sup_{\mu \in P(X)} \int f d\mu - D_{KL}(\mu \| \nu) = \log \int e^f d\nu$$

Biconjugate representation leads to the Donsker-Varadhan formula:

$$D_{KL}(\mu \| \nu) = D^{**}_{KL}(\mu \| \nu) = \sup_{f : X \to \mathbb{R}} \int f d\mu - \log \int e^f d\nu$$

Csiszár potentials

$$D_{KL}(\mu \| \nu) = \int f_* d\mu - \log \int e^{f_*} d\nu$$

if and only if $\mu \ll \nu$ and $\exists C \in \mathbb{R}$ such that

$$f_* = \log \frac{d\mu}{d\nu} + C$$

For the binary experiment of discriminating samples from μ and ν, the statistical test $(x \to \chi_{[\tau,\infty]}(f_*(x)))$ is a most powerful test for any threshold $\tau \in \mathbb{R}$.

GAN perspective

Tight f-GAN: optimal critic is not Lipschitz continuous in general!

Wasserstein-1 distance and Kantorovich potentials

$\mu, \nu \in P(X)$ for a metric space (X, d):

$$W_1(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int d(x_1, x_2) d\pi(x_1, x_2)$$

$$= \sup_{f: X \to \mathbb{R}, \|f\|_L \leq 1} \int f d\mu - \int f d\nu$$

$$W_1(\mu, \nu) = \int d(x_1, x_2) d\pi_*(x_1, x_2) = \int f_* d\mu - \int f_* d\nu$$

if and only if

$$\|f_*\|_L = 1$$

$$f_*(x_1) - f_*(x_2) = d(x_1, x_2) \ \pi_*\text{-a.e.}$$

GAN perspective

Wasserstein GAN: Lipschitz constant of optimal critic 1

Unconstrained Kantorovich-Rubinstein formula

$$\lambda W_1(\mu,\nu)^\alpha = \sup_{f:X\to\mathbb{R}} \int f d\mu - \int f d\nu - (\alpha-1)\alpha^{\frac{\alpha}{1-\alpha}}\lambda^{\frac{1}{1-\alpha}}\|f\|_L^{\frac{\alpha}{\alpha-1}}$$

$$\lambda W_1(\mu,\nu)^\alpha = \int f_* d\mu - \int f_* d\nu - (\alpha-1)\alpha^{\frac{\alpha}{1-\alpha}}\lambda^{\frac{1}{1-\alpha}}\|f_*\|_L^{\frac{\alpha}{\alpha-1}}$$

if and only if

$$\frac{1}{\alpha\lambda W_1(\mu,\nu)^{\alpha-1}} f_*$$

is a Kantorovich potential of μ, ν.

GAN perspective

Unconstrained Wasserstein GAN: Lipschitz constant of optimal critic is $\alpha\lambda W_1(\mu,\nu)^{\alpha-1}$

Moreau-Yosida approximation of D_{KL} with respect to W_1

$$
\begin{aligned}
D_{KL,\lambda,\alpha}(\mu\|\nu) &= \min_{\xi\in P(X)} D_{KL}(\xi\|\nu) + \lambda W_1(\mu,\xi)^\alpha \\
&= \sup_{f:X\to\mathbb{R}, \|f\|_L\leq\lambda} \int f d\mu - \log \int e^f d\nu \text{ if } \alpha = 1 \\
&= \sup_{f:X\to\mathbb{R}} \int f d\mu - \log \int e^f d\nu - (\alpha-1)\alpha^{\frac{\alpha}{1-\alpha}} \lambda^{\frac{1}{1-\alpha}} \|f\|_L^{\frac{\alpha}{\alpha-1}} \text{ if } \alpha > 1
\end{aligned}
$$

Csiszár-Kantorovich potentials

$$D_{KL,\lambda,\alpha}(\mu\|\nu) = D_{KL}(\xi_*\|\nu) + \lambda W_1(\mu,\xi_*)^\alpha$$
$$= \int f_* d\mu - \log \int e^{f_*} d\nu - (\alpha-1)\alpha^{\frac{\alpha}{1-\alpha}}\lambda^{\frac{1}{1-\alpha}}\|f_*\|_L^{\frac{\alpha}{\alpha-1}}$$

if and only if $\xi_* \ll \nu$ and $\exists C \in \mathbb{R}$ such that $f_* = \log\frac{d\xi_*}{d\nu} + C$ and $\frac{1}{\alpha\lambda W_1(\mu,\xi_*)^{\alpha-1}} f_*$ is a Kantorovich potential of μ, ξ_*, so that $\|f_*\|_L = \alpha\lambda W_1(\mu,\xi_*)^{\alpha-1}$.

GAN perspective

Moreau-Yosida f-GAN: Lipschitz constant of optimal critic is $\alpha\lambda W_1(\mu,\xi_*)^{\alpha-1}$

That's not all folks!

See the paper for

- generalization to f-divergences and
- GAN experiments.