Infotheory for Statistics and Learning

Lecture 7

- Donsker–Varadhan [PW:4.3]
- Variational characterization of f-divergence [PW:7.13]
- Marginalization and the ELBO [MK:33]
- Variational free energy and inference [MK:33]

Donsker–Varadhan

For P and Q on $(\Omega, \mathcal{A})$ we have

$$D(P\|Q) = \sup_X \left\{ \int X(\omega) dP - \ln \int e^{X(\omega)} dQ \right\}$$

(for $D(P\|Q)$ in nats) where the supremum is over RVs X such that $E_Q[\exp(X(\omega))] < \infty$

Proof: Let $Y(X) = E_P[X]$ and $Z(X) = \ln E_Q[\exp(X)]$, then

$$D(P\|Q) = Y\left(\ln\frac{dP}{dQ}\right) - Z\left(\ln\frac{dP}{dQ}\right)$$

Also, for $A \in \mathcal{A}$ define

$$Q_X(A) = \int_A \exp(X(\omega) - Z(X))dQ$$

then

$$Y(X) - Z(X) = E_P\left[\ln\left(\frac{dP}{dQ}\frac{dQ_X}{dP}\right)\right] = D(P\|Q) - D(P\|Q_X)$$

f-divergence

Remember the Fenchel–Legendre dual, given a (convex) function $f(x)$, define

$$f^*(y) = \sup_x(xy - f(x))$$

Then (assuming $P \ll Q$)

$$D_f(P\|Q) = E_Q\left[f\left(\frac{dP}{dQ}(\omega)\right)\right] = \sup_X\left\{E_P[X] - E_Q[f^*(X)]\right\}$$

and the supremum is obtained for

$$X(\omega) = f'\left(\frac{dP}{dQ}(\omega)\right)$$

Proof: For fixed Q, $D_f(P\|Q)$ is convex in P. Let

$$D_f^*(X) = \sup_P \left\{ E_P[X] - E_Q\left[f\left(\frac{dP}{dQ}\right)\right]\right\}$$

be the dual of $P \mapsto D_f(P\|Q)$. Then since $(D_f^*)^* = D_f$

$$D_f(P\|Q) = \sup_X \left\{ E_P[X] - D_f^*(X)\right\}$$

Also, since $f^*(X) = \sup_\omega(\omega X - f(\omega))$ we have

$$f^*(X(\omega)) \geq X(\omega)\frac{dP}{dQ}(\omega) - f\left(\frac{dP}{dQ}(\omega)\right)$$

$$\Rightarrow E_Q[f^*(X)] \geq E_P[X] - D_f(P\|Q) \Rightarrow E_Q[f^*(X)] \geq D_f^*(X)$$

Hence

$$D_f(P\|Q) \geq \sup_X \{E_P[X] - E_Q[f^*(X)]\}$$

On the other hand, for

$$X(\omega) = f'\left(\frac{dP}{dQ}(\omega)\right)$$

we get

$$f^*(X) = \frac{dP}{dQ} f'\left(\frac{dP}{dQ}\right) - f\left(\frac{dP}{dQ}\right)$$

since $f^*(f'(x)) = xf'(x) - f(x)$, that is

$$D_f(P\|Q) = E_P\left[f'\left(\frac{dP}{dQ}(\omega)\right)\right] - E_Q\left[f^*\left(f'\left(\frac{dP}{dQ}(\omega)\right)\right)\right]$$

Special cases

With $f(x) = x \ln x$ we get $f^*(y) = \exp(y-1)$ and

$$D_f(P\|Q) = D(P\|Q) = \sup_X \{E_P[X] - E_Q[\exp(X-1)]\}$$

to compare with Donsker–Varadhan

$$D(P\|Q) = \sup_X \{E_P[X] - \ln E_Q[\exp(X)]\}$$

Since $\ln t \le t/e$ the lower bound obtained from D–V is tighter

With $f(x) = (x-1)^2$ we get $f^*(y) = y + y^2/4$ and

$$D_f(P\|Q) = \chi^2(P\|Q) = \sup_X \left\{ E_P[X] - E_Q\left[X + \frac{X^2}{4}\right]\right\}$$

Or, setting $Y(\omega) = X(\omega)/2 + 1$,

$$\chi^2(P\|Q) = \sup_Y \left\{2E_P[Y] - E_Q\left[Y^2\right] - 1\right\}$$

Marginalization and the ELBO

Assume $X^n \in \mathbb{R}$ and $Y^m \in \mathbb{R}$ are jointly distributed according to a pdf $p_\theta(x^n, y^m)$ where $\theta \in \mathbb{R}^d$ is a parameterization

Consider e.g. the ML problem based on observing $X^n = x^n$ but not the *latent variables* Y^m, where we want to compute

$$p_\theta(x^n) = \int p_\theta(x^n, y^m) dy^m$$

(or the corresponding sum if y^m is discrete)

This is a typical marginalization problem, which can be hard or impossible to solve if $m \gg 1$

Let $q(y^m|x^n)$ be an arbitrary conditional pdf for y^m given x^n, chosen from a class $\mathcal{Q}$

Then define the evidence lower bound (ELBO) as

$$\mathcal{L}_\theta(x^n; q) = \ln p_\theta(x^n) - D(q(y^m|x^n) \| p_\theta(y^m|x^n))$$

where $p_\theta(y^n|x^n)$ is the true pdf and the divergence is over y^m

That is, if $\mathcal{Q}$ contains $p_\theta(y^m|x^n)$ then

$$\ln p_\theta(x^n) = \max_{q \in \mathcal{Q}} \mathcal{L}_\theta(x^n; q)$$

$\Rightarrow$ *marginalization by optimization*

But can $\mathcal{L}_\theta(x^n; q)$ be computed?

Connection to statistical physics

In physics one often assumes canonical models of the form

$$p_\theta(x^n) = \frac{1}{Z(\theta)} \exp\left(-\beta E(x^n;\theta)\right)$$

where $p_\theta(x^n)$ is the pmf for $x^n \in \{\pm 1\}^n$, with ± 1 corresponding e.g. to "spin up" or "spin down"

Also, $E(x^n;\theta)$ is the energy function and

$$Z(\theta) = \sum_{x^n} e^{-\beta E(x^n;\theta)}$$

is the partition function

The variational free energy for the system is

$$\beta\tilde{F}(\theta) = \sum_{x^n} q_\theta(x^n) \ln \frac{q_\theta(x^n)}{\exp(-\beta E(x^n;\theta))}$$

relative to the pmf $q_\theta(x^n)$

Note that

$$\begin{aligned} \beta\tilde{F}(\theta) &= \beta \sum_{x^n} q_\theta(x^n) E(x^n;\theta) - H(q_\theta) \\ &= \beta F(\theta) + D(q_\theta(x^n) \| p_\theta(x^n)) \end{aligned}$$

where $\beta F(\theta) = -\ln Z(\theta)$ is the true free energy

$\Rightarrow$ Approximate $\beta F(\theta)$ by choosing q_θ to minimize $\beta\tilde{F}(\theta)$

Note the relation to the ELBO: We marginalize over x^n to get $\beta F(\theta)$, so ELBO $\leftrightarrow$ negative of the variational free energy

Mean field equations for a spin system

For the energy function

$$E(x^n;\theta,h) = -\frac{1}{2}\sum_{ij}\theta_{ij}x_ix_j - \sum_i h_ix_i$$

use a $q_\theta(x^n)$ of the form

$$q_a(x^n) = \frac{\exp\left(\sum_i a_ix_i\right)}{\sum_{x^n}\exp\left(\sum_i a_ix_i\right)}$$

$\Rightarrow$ *mean field equations* to minimize $\beta\tilde{F}$

$$a_i = \beta\sum_j \theta_{ij}\bar{x}_j + \beta h_i$$
$$\bar{x}_i = \tanh(a_i)$$

Connection to the dual

Consider $E(x^n;\theta) = -\sum\theta_i\phi_i(x_i) = -\langle\theta,\phi(x^n)\rangle$ so that

$$p_\theta(x^n) = \exp\left(\beta\langle\theta,\phi(x^n)\rangle + \beta F(\theta)\right)$$

Let $\beta F^*(\mu) = \sup_\theta(\beta F(\theta) + \langle\theta,\mu\rangle)$

The sup is achieved when the relation $\mu(\theta) = \beta E_\theta[\phi(X^n)]$ holds

Note that $E_\theta[\ln p_\theta(X^n)] = \beta\langle\theta, E_\theta[\phi(X^n)]\rangle + \beta F(\theta) = -H(p_\theta)$

$\Rightarrow \beta F^*(\mu) = -H(p_\theta)$ for θ such that $\mu = \beta E_\theta[\phi(X^n)]$

For a general μ (not coupled to θ), $\beta F(\theta) \leq \beta F^*(\mu) - \langle\theta,\mu\rangle$

For any pmf $q(x^n)$ we also have (due to Jensen)

$$\beta F(\theta) \leq \sum_{x^n} q(x^n)\ln q(x^n) - \beta\sum_{x^n} q(x^n)\langle\theta,\phi(x^n)\rangle = \beta\tilde{F}(\theta,q)$$

General mean field problem:
Compute/approximate $\mu^* = \beta E_\theta[\phi(X^n)]$ by minimizing $\beta\tilde{F}(\theta,q)$ over q to get p_θ or $\beta F^*(\mu) - \langle\theta,\mu\rangle$ directly over μ

Other choices for q_θ

Alternative *separable* models for q_θ of the form $q_\theta(x^n) = \prod q_{\theta_i}(x_i)$, or a more general *Markov structure* described by a *factor graph*

Alternative physics-based methods, such as Bethe or Kikuchi free energy models

See [MK] and

Wainwright & Jordan, "Graphical models, exponential families and variational inference," *FnT's Machine Learning 2008*

for a thorough account

Variational Bayes

For $\theta \in \Theta$ and data X consider the average and Bayes risks

$$R_\pi(\hat{\theta}) = \int \left\{ \int \ell(\theta, \hat{\theta}(x)) P_\theta(dx) \right\} \pi(d\theta)$$

$$R_\pi^* = \inf_{\hat{\theta}} R_\pi(\hat{\theta})$$

Let $\hat{\theta}^*$ denote the corresponding Bayes estimator

For $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ we get

$$\hat{\theta}^*(x) = \int \theta dP_{\theta|X=x}$$

and for $|\Theta| < \infty$ and $\ell(\theta, \hat{\theta}) = \mathbb{1}(\theta \neq \hat{\theta})$ we get

$$\hat{\theta}^*(x) = \arg\max_\theta p_{\theta|X=x}(\theta|x)$$

Thus for Bayesian inference in general, we need access to the conditional or posterior distribution $P_{\theta|X=x}$, either via a pdf $f_{\theta|X=x}(\theta|x)$ or a pmf $p_{\theta|X=x}(\theta|x)$

We have (for the case of a pdf)

$$f_{\theta|X=x}(\theta|x) = \frac{f_\theta(x)\pi(\theta)}{\int f_\theta(x)\pi(\theta)d\theta}$$

where we have a marginalization problem in computing the integral

$$f(x) = \int f_\theta(x)\pi(\theta)d\theta$$

With

$$\mathcal{L}(q) = -\int q(\theta)\ln\frac{q(\theta)}{f_\theta(x)\pi(\theta)}d\theta = \ln f(x) - D(q\|f_{\theta|X=x})$$

we can maximize over q to compute/approximate $\ln f(x)$

Example: Assume $X^n = (X_1, \dots, X_n)$ drawn iid $\sim \mathcal{N}(\mu, \sigma^2)$ and consider $\theta = (\mu, \sigma)$ for the model

$$f_{\theta|X^n=x^n}(\theta|x^n)f(x^n) = \frac{1}{\sigma_\mu\sigma(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{n(\mu-\bar{x})^2+S}{2\sigma^2}\right)$$

(corresponding to an *improper noninformative prior* for μ and σ, see MK Ch. 24) where $\bar{x} = n^{-1}\sum_i x_i$ and $S = \sum_i(x_i-\bar{x})^2$

To compute $f(x^n)$ we seek to minimize

$$\int q(\theta)\ln\frac{q(\theta)}{f_{\theta|X^n=x^n}(\theta|x^n)f(x^n)}d\theta$$

over $q(\theta)$ of the form $q(\theta) = q(\mu)q(\sigma)$ (separable)

Thus, for a fixed $q(\sigma)$ we minimize

$$\int q(\mu)\left\{\int q(\sigma)\frac{n(\mu-\bar{x})^2}{2\sigma^2}d\sigma+\ln q(\mu)\right\}d\mu$$

$\Rightarrow q(\mu)=\mathcal{N}(\bar{x},\sigma^2_{\mu|x^n})$ where

$$\sigma^2_{\mu|x^n}=\frac{1}{n\int q(\sigma)/\sigma^2 d\sigma}$$

Similarly, for a fixed $q(\mu)$ we get

$$q(\beta)=\Gamma(\beta;a,b)=\frac{1}{a\Gamma(b)}\left(\frac{\beta}{a}\right)^{b-1}e^{-\beta/a}$$

with $\beta=1/\sigma^2$ and where $1/a=(n\sigma^2_{\mu|x^n}+S(x^n))/2$ and $b=n/2$

Since $\int\beta q(\beta)d\beta=ab$ we also have $\sigma^2_{\mu|x^n}=S(x^n)/(n(n-1))$