IFT 6269: Probabilistic Graphical Models **Fall 2018**

Lecture 16 — November 2

Lecturer: Simon Lacoste-Julien *Scribe: Tapopriya Majumdar*

Disclaimer: Lightly proofread and quickly corrected by Simon Lacoste-Julien.

16.1 Information Theory

16.1.1 Kullback–Leibler (KL) Divergence

For discrete distributions p and q, the KL divergence between p and q is defined to be

$$D_{\mathrm{KL}}(p \,||\, q) \triangleq \sum_{x \in \Omega} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p \left[\log \frac{p(x)}{q(x)} \right] \tag{16.1}$$

Motivation from density estimation

Let $\hat{q}$ be an estimation of the given distribution. Recall the statistical decision theory setting. The standard (Maximal Likelihood) loss is the log-loss, giving the following statistical loss when the true distribution is p_θ for action $\hat{q}$:

$$L(p_\theta, \hat{q}) \triangleq \mathbb{E}_{X \sim p_\theta}[-\log \hat{q}(X)] \tag{16.2}$$

Note that above is called the cross-entropy. If we use the best action $\hat{q} = p_\theta$, then we get the loss to be

$$-\sum_{x \in \Omega} p_\theta(x) \log \; p_\theta(x) = H(p_\theta), \tag{16.3}$$

the entropy of p_θ (which is obviously the best we can do, as we are outputting the correct distribution). Therefore, the excess loss in this case is

$$\begin{aligned} L(p, \hat{q}) - \min_q L(p, q) &= L(p, \hat{q}) - L(p, p) \\ &= -\sum_{x \in \Omega} p(x) \log \frac{\hat{q}(x)}{p(x)} \\ &= D_{\mathrm{KL}}(p \,||\, \hat{q}) \end{aligned}$$

So the KL divergence can be interpreted as the excess log-loss we get by outputting $\hat{q}$ instead of the true distribution p.

Motivation from coding theory

We use the fact that in coding theory, the optimal length of a code is proportional to $-\log_2 p(x)$ bits. Then the expected length of the code is $\sum_x p(x)(-\log_2 p(x))$, where the entropy is measured in bits.[1] Then the KL divergence can be interpreted as the excess cost (in terms of length of code) to use a distribution q for coding as opposed to the optimal distribution p.

16.1.2 Examples

Example 16.1.1 (Entropy of a Bernoulli distribution)
Let $X \sim \mathbf{B}ern(p)$. Then

$$H(X) = -p \log p - (1-p)\log(1-p), \tag{16.4}$$

which is largest when $p = 1/2$.

Example 16.1.2 (Entropy of a uniform distribution on K states)
Let $X \sim \mathrm{U}niform(\{x_1, \ldots, x_K\}$. Then

$$H(X) = -\sum_{i=1}^{K} \frac{1}{K} \log \frac{1}{K} = \log K \tag{16.5}$$

It turns out that the uniform distribution on K states is the one with maximum entropy, among all distributions over K states.

16.1.3 Properties

1. $D_{\mathrm{KL}}(p \,||\, q) \geqslant 0$. This can be shown using Jensen's equality.
2. It is strictly convex in each argument.
3. It is not symmetric: $D_{\mathrm{KL}}(p \,||\, q) \neq D_{\mathrm{KL}}(q \,||\, p)$.
4. $D_{\mathrm{KL}}(p \,||\, p) = 0 \; \forall p$ and $D_{\mathrm{KL}}(p \,||\, q) > 0$ when $p \neq q$.

16.1.4 Maximal Likelihood and KL Minimization

Let $\{p_\theta\}_{\theta \in \Theta}$ be a parametric family of distributions, and $\hat{p}_n(x) \triangleq \frac{1}{n} \sum_{i=1}^{n} \delta\left(x, x^{(i)}\right)$ be the empirical distribution corresponding to n samples. Then

$$\text{ML for } \theta \iff \min_{\theta \in \Theta} D_{\mathrm{KL}}(\hat{p}_n \,||\, p_\theta). \tag{16.6}$$

[1] When using log in the natural base, the entropy is measured in nats, when using $\log_2$, it is measured in bits.

Proof

$$\begin{aligned} D_{\mathrm{KL}}(\hat{p}_n \,||\, p_\theta) &= \sum_{x\in\Omega} \hat{p}_n(x) \log \frac{\hat{p}_n(x)}{p_\theta(x)} \\ &= H(\hat{p}_n) - \sum_{x\in\Omega} \hat{p}_n \log p_\theta(x) \\ &= H(\hat{p}_n) - \frac{1}{n} \sum_{x\in\Omega} \sum_{i=1}^{n} \delta\left(x, x^{(i)}\right) \log p_\theta(x) \\ &= H(\hat{p}_n) - \frac{1}{n} \sum_{i=1}^{n} \log p_\theta\left(x^{(i)}\right) \\ &= \text{constant} - \frac{1}{n} \log \prod_{i=1}^{n} p_\theta\left(x^{(i)}\right). \end{aligned}$$

16.2 Maximum Entropy Principle

Here the idea is to consider some subset of distributions over $\mathcal{X}$ according to some data-driven constraint, i.e. a subset $\mathcal{M} \subseteq \Delta_{|\mathcal{X}|}$. The principle is to pick $\hat{p} \in \mathcal{M}$ which maximizes the entropy:

$$\begin{aligned} \hat{p} &= \text{argmax}_{q\in\mathcal{M}} H(q) \\ &= \text{argmin}_{q\in\mathcal{M}} D_{\mathrm{KL}}(q \,||\, \text{uniform}), \end{aligned}$$

as $D_{\mathrm{KL}}(q \,||\, \text{uniform}) = -H(q) + \text{constant}$.

More generally, we can also consider the generalized maximum entropy principle where we do: $\arg\min_{q\in\mathcal{M}} D_{\mathrm{KL}}(q \,||\, h_0)$, for some distribution h_0 that we want to favor (instead of the uniform, which is used for the standard maximum entropy). We'll see soon the role of this h_0 when we talk about the equivalence of maximum entropy with maximum likelihood in the exponential family.

Example 16.2.1 (from Wainwright) *If we observe $p_L = 3/4$ kangaroos are left-handed and $p_B = 2/3$ kangaroos drink Labatt beer, then how many kangaroos are both left-handed and drink Labatt beer? (Here the max. entropy solution is that $p(B, L) = p_B \cdot p_L$, by independence)*

16.2.1 How do we get $\mathcal{M}$?

A standard way to get $\mathcal{M}$ is through empirical "moments": let the feature functions be $T_1(x), \ldots, T_d(x)$ – the represent various measurements we want to make on the data. Then define $\mathcal{M} \triangleq \{q : \mathbb{E}_q[T_j(x)] = \mathbb{E}_{\hat{p}_n}[T_j(x)] \,\forall\, j = 1, \ldots, d\}$, that is, the set of distributions for which their model moments match the empirical moments. If we let $\alpha_j \triangleq \mathbb{E}_{\hat{p}_n}[T_j(x)]$. Then the constraint becomes $\sum_x q(x) T_j(x) = \alpha_j$ (some scalar), i.e. $\langle q, T_j \rangle = \alpha_j$ (it's a linear

equality on q, when it is represented as a vector over $|\mathcal{X}|$ elements). Hence, finding q using Maximal Entropy

$$\min_{q \in \mathbb{R}^{|\mathcal{X}|}} D_{\mathrm{KL}}(q \,||\, \text{uniform}) \text{ such that } q \in \mathcal{M} \cap \Delta_{|\mathcal{X}|}$$

becomes a convex optimization problem over $q \in \Delta_{|\mathcal{X}|} \subseteq \mathbb{R}^{|\mathcal{X}|}$.

16.2.2 Lagrangian duality segue

Let $f, f_j, j = 1, \ldots, m$ be convex functions and $g_k, k = 1, \ldots, n$ be affine functions. Here these functions are extended real-valued functions, e.g. $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$. then $\mathrm{dom}(f) \triangleq \{x : f(x) < \infty\}$. The primal convex optimization problem is:

$$\begin{aligned} \text{minimize}_x f(x) & \\ \text{such that } f_j(x) &\leqslant 0 \,\forall j \\ \text{and } g_k(x) &= 0 \,\forall k \end{aligned}$$

We define

$$\mathcal{L}(x, \lambda, \nu) \triangleq f(x) + \sum_{j=1}^{m} \lambda_j f_j(x) + \sum_{k=1}^{n} \nu_k g_k(x),$$

where λ_j and ν_k are Langrange multipliers. We will now present the saddle point interpretation of the Lagrangian duality. It uses the following trick:

$$h(x) \triangleq \sup_{\substack{\lambda \geqslant 0 \\ \nu}} f(x, \lambda, \nu) = \begin{cases} f(x) & \text{if } x \text{ is feasible} \\ +\infty & \text{if } x \text{ is not feasible} \end{cases} \tag{16.7}$$

so an equivalent problem to the (constrained) primal problem is the following (unconstrained) problem using the fancy complicated function $h(x)$:

$$\inf_x \Bigg(\underbrace{\sup_{\substack{\lambda \geqslant 0 \\ \nu}} f(x, \lambda, \nu)}_{h(x)} \Bigg). \tag{16.8}$$

The duality trick is to swap inf and sup:

$$\sup_{\substack{\lambda \geqslant 0 \\ \nu}} \left(\inf_x f(x, \lambda, \nu) \right). \tag{16.9}$$

Lagrangian dual problem

Let $\inf_x f(x, \lambda, \nu) \triangleq g(\lambda, \nu)$, so that g is always concave in both components. The Lagrangian dual problem is to solve

$$\sup_{\substack{\lambda \geqslant 0 \\ \nu}} g(\lambda, \nu). \tag{16.10}$$

The weak duality

$$\sup_{\substack{\lambda \geqslant 0 \\ \nu}} \inf_x f(x, \lambda, \nu) \leqslant \inf_x \sup_{\substack{\lambda \geqslant 0 \\ \nu}} f(x, \lambda, \nu)$$

is always true (because $\sup \inf \leq \inf \sup$ always). Let $p^* \triangleq \{\inf_x f(x) : x \text{ feasible}\}$. Then $g(\lambda, \nu) \leqslant p^* \, \forall \, \lambda \geqslant 0, \nu$. The strong duality is when we have equality, i.e.

$$d^* = \sup_{\substack{\lambda \geqslant 0 \\ \nu}} g(\lambda, \nu) = p^*. \tag{16.11}$$

When the primal optimization problem is convex, a sufficient condition for strong duality is Slater's condition:[2] $\exists \, x \in \text{int}(\text{dom}(f))$ such that $f_j(x) < 0 \, \forall \, j$ where f_j is nonlinear and x is feasible. See the Chapter 5 in Boyd's book http://stanford.edu/~boyd/cvxbook/ for more details.

Note that after solving the dual problem and obtaining λ^*, ν^*, one can usually reconstruct the primal optimal variables $x^*(\lambda^*, \nu^*)$ (when strong duality holds) using the KKT conditions, which are a set of necessary non-linear equations that hold for the primal and dual optimal variables.

16.3 Dual Problem for Maximal Entropy

Let u be the uniform distribution on $\mathcal{X}$. Let $\Delta_{|\mathcal{X}|} \triangleq \{q : q(x) \geqslant 0 \, \forall \, x, \sum_x q(x) = 1\}$ and $\mathcal{M} \triangleq \{q \in \Delta_{|\mathcal{X}|} : \sum_x q(x) T_j(x) = \alpha_j \, \forall \, j\}$. Then the primal form of the maximal entropy problem is to find

$$\min_{q \in \mathcal{M}} \sum_x q(x) \log \frac{q(x)}{u(x)} \tag{16.12}$$

As we did in the lecture on deriving the maximum likelihood parameter for the multinouilli, we will ignore the inequality constraints on q ($q(x) \geq 0$), as the KL divergence is essentially acting as a barrier function making sure that q stays positive. So we only form the Lagrangian with ν for the moment equality constraints, and we use a separate Lagrange multiplier c for the sum-to-one equality constraint, as we'll see later that we will treat it differently.

We thus introduce the corresponding Lagrangian

$$\mathcal{L}(q, \nu, c) \quad = \quad \sum_x q(x) \log \frac{q(x)}{u(x)} + \sum_j \nu_j \left(\alpha_j - \mathbb{E}_q[T_j(x)]\right) + c \left(1 - \sum_x q(x)\right)$$

To get the dual function, we need to minimize the Lagrangian with respect to q (it is convex in q, so we just need to find its zero gradient):

$$\text{We have,} \qquad \frac{\partial \mathcal{L}}{\partial q(x)} \quad = \quad 1 + \log \frac{q(x)}{u(x)} - \sum_j \nu_j T_j(x) - c$$

[2] This is an example of constraint qualification condition; there are others.

So

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial q(x)} = 0 \quad &\iff \quad \log \frac{q^*(x)}{u(x)} = \langle \nu, T \rangle + c - 1 \\ &\iff \quad q^*_{\nu,c}(x) = u(x) \exp\left(\langle \nu, T \rangle + c - 1\right),\end{aligned}$$

so that q^* is part of the exponential family of distributions!

Dual Function

Plugging in this value of q^* in $\mathcal{L}$, (we use the (abused) shorthand notation $\mathbb{E}_{q^*}$ below to denote $\sum_x q^*(x)$ even though q^* is not necessarily normalized):

$$\begin{aligned} g(\nu, c) &= \mathcal{L}(q^*_{\nu,c}, \nu, c) \\ &= \mathbb{E}_{q^*}[\langle \nu, T(x) \rangle + c - 1] + \langle \nu, \alpha \rangle - \mathbb{E}_{q^*}[\langle \nu, T(x) \rangle] + c - \mathbb{E}_{q^*}[c] \\ &= \langle \nu, \alpha \rangle + c - \mathbb{E}_{q^*}[1] \\ &= \langle \nu, \alpha \rangle + c - \sum_x u(x) \exp\left(\langle \nu, T(x) \rangle\right) \exp(c - 1) \\ &= \langle \nu, \alpha \rangle + c - Z(\nu) \exp(c - 1), \end{aligned}$$

where $Z(\nu) \triangleq \sum_x u(x) \exp\left(\langle \nu, T(x) \rangle\right)$. Therefore,

$$\frac{\partial g}{\partial c} = 1 - Z(\nu) \exp(c - 1).$$

To maximize $g(\nu, c)$ with respect to c,

$$\begin{aligned}\frac{\partial g}{\partial c} = 0 \quad &\iff \quad 1 - Z(\nu) \exp(c^* - 1) = 0 \\ &\iff \quad \exp(c^* - 1) = \frac{1}{Z(\nu)}\end{aligned}$$

Plugging back c^*, we get

$$\begin{aligned} \max_c g(\nu, c) &= \langle \nu, \alpha \rangle + c^* - Z(\nu) \exp(c^* - 1) \\ &= \langle \nu, \alpha \rangle + c^* - Z(\nu) \frac{1}{Z(\nu)} \\ &= \langle \nu, \alpha \rangle + c^* - 1 \\ &= \langle \nu, \alpha \rangle + \log Z(\nu) \\ &\triangleq \tilde{g}(\nu) \end{aligned}$$

By eliminating c from the dual problem, we ensure that q^*_{ν,c^*} is normalized (which is why we treated it differently). $\tilde{g}$ is the corresponding objective for the remaining dual problem. We

now re-interpret this dual problem and link it with maximum likelihood for the exponential family.

If $\alpha = \frac{1}{n}\sum_i T\left(x^{(i)}\right) = \mathbb{E}_{\hat{p}_n}[T(x)]$, then

$$\begin{aligned}\tilde{g}(\nu) &= \frac{1}{n}\sum_{i=1}^{n}\left[\langle \nu, T\left(x^{(i)}\right)\rangle - \log(Z(\nu)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\log p\left(x^{(i)}|\nu\right),\end{aligned}$$

where $p(x|\nu) \triangleq u(x)\exp\left(\langle \nu, T(x)\rangle - \log Z(\nu)\right)$. Then the dual problem is

$$\max_{\nu}\tilde{g}(\nu) = \max_{\nu}\frac{1}{n}\log p\left(x^{(1):(n)}|\nu\right),$$

which is the same as the maximal likelihood estimate!

To summarize, maximal likelihood in the exponential family with $T(x)$ as the sufficient statistics is equivalent to the maximal entropy problem with moment constraints on $T(x)$, where $\alpha = \mathbb{E}_{\hat{p}_n}[T(x)]$. They are Lagrangian dual of one another:

MLE in exponential family $\iff$ maximum entropy with moment constraints

Note moreover that if we use the generalized maximum entropy principle $\arg\min_{q\in\mathcal{M}} D_{\mathrm{KL}}(q\,||\,h_0)$ with h_0 instead of the uniform, then we get an exponential family with $h_0(x)$ as the reference density instead of the uniform distribution!

Remark 16.3.1

$$\begin{aligned}\nabla_\nu \log Z(\nu) &= \frac{1}{Z(\nu)}\nabla_\nu \sum_x u(x)\exp(\langle \nu, T(x)\rangle) \\ &= \sum_x \frac{1}{Z(\nu)}T(x)u(x)\exp(\langle \nu, T(x)\rangle) \\ &= \sum_x p(x|\nu)T(x) \\ &= \mathbb{E}_{p(x|\nu)}[T(x)] \\ &\triangleq \mu(\nu),\ \textit{the ``model moment''}\end{aligned}$$

Therefore,

$$\begin{aligned}\nabla_\nu\tilde{g}(\nu) &= \mathbb{E}_{\hat{p}_n}[T(x)] - \mu(\nu) \\ &\triangleq \hat{\mu}_n - \mu(\nu),\end{aligned}$$

where $\hat{\mu}_n$ is the "empirical moment". We note that

$$\nabla_\nu\tilde{g}(\nu) = 0 \Rightarrow \mu(\nu^*) = \hat{\mu}_n,$$

i.e. the maximal likelihood parameters in the exponential family are also doing moment matching (which is expected by the equivalence above).

So in the case of the exponential family, we have that maximum likelihood is equivalent to maximum entropy which is equivalent to moment matching. For other parametric families (mixture models for example, which are not in the exponential family), then moment matching could give a different estimator than maximum likelihood.