Lecture 19 - MaxEnt duality; exp. family

- max Ent & duality
- exponential family

MLE & KL minimization

\[ \text{MLE} \Rightarrow \text{min } \text{KL} \left( \hat{p}_n \parallel p_0 \right) \]

Proof:
\[
\text{KL} \left( \hat{p}_n \parallel p_0 \right) = \sum \hat{p}_n(x) \log \frac{\hat{p}_n(x)}{\hat{p}_0(x)} \\
= -H(\hat{p}_n) - \sum \hat{p}_n(x) \log \frac{\hat{p}_0(x)}{\hat{p}_0(x)} \\
= -H(\hat{p}_n) - \frac{1}{n} \sum \log \frac{\hat{p}_0(x)}{\hat{p}_0(x)} \\
= \text{constant w.r.t. } p_0 \]

Maximum entropy principle:

Idea: consider some subset of dist. on \( X \) according to some data-driven constraints

get a subset \( M \subseteq \Delta |X| \) - pick simple \( |X| = k \) elements

MAXENT principle: pick \( \hat{p} \in M \) which maximizes the entropy

\[ \hat{p} = \arg \max_{q \in M} H(q) \]

\[ \hat{p} = \arg \max_{q \in M} \text{KL}(q \parallel \text{uniform}) \]

\[ \text{KL}(q \parallel \text{uniform}) = \sum q(x) \log \frac{q(x)}{1/|X|} = -H(q) + \text{const} \]

"generalized max entropy" \( \text{KL}(q \parallel p_0) \)

preferred dist. to \( p_0 \) has max entropy
Example from Wainwright:
\[ \hat{p}_c = \frac{3}{4} \text{ kangaroo is left-handed} \]
\[ \hat{p}_B = \frac{2}{3} \text{ drink talent here} \]

Question: How many kangaroos drink soda here?

[Note: max. entropy solution is that \( p(B|E) = \hat{p}_B \cdot p_e \) (indep.)]

How do we get set \( M \)?

- Typically: through empirical "moments"

  Feature functions: \( T_1(x), \ldots, T_d(x) \)
  Define \( M = \{ q \in \mathbb{R}^d : \mathbb{E} q(T_j(x)) = \mathbb{E}_{\hat{p}_B} q(T_j(x)) \} \)

  \( \mathbb{E} q(T_j(x)) \) is empirical feature count
  \( \mathbb{E}_{\hat{p}_B} q(T_j(x)) \) is model expected feature count

  "moment constraints"

\[ \min_{q \in M} \max_{\lambda, \mu} \mathbb{L}(q \parallel \mu) \text{ s.t. } \int q(x) \, dx = 1 \]

\[ \mathbb{L}(q \parallel \mu) = \int q(x) \log \frac{q(x)}{\mu(x)} \, dx \]

\[ \mathbb{L}(q \parallel \mu) = \int q(x) \log \frac{q(x)}{\mu(x)} \, dx \]

Quick presentation of Lagrangean duality:

Convex min problem: \( \min_{x \in \mathbb{R}^n} f(x) \)

Equivalent problem \( \{ f(x), f_i(x), f_k(x) \text{ s.t. } f_i(x) \leq 0 \text{ for all } i \} \)

Lagrangean function: \( \mathbb{L}(x, \lambda, \mu) \)

\[ \mathbb{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{k=1}^n \mu_k f_k(x) \]

Magic trick (saddle pt. interpretation):

\[ b(x) = \sup_{\lambda, \mu} \mathbb{L}(x, \lambda, \mu) \]

\[ b(x) = \sup_{\lambda, \mu} \left\{ f(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{k=1}^n \mu_k f_k(x) \right\} \]

An equivalent problem to the primal problem...
see chapter 5 in Boyd's book for more info on duality: http://stanford.edu/~boyd/cvxbook/

**Dual problem for max entropy**

\[
\begin{align*}
\text{MaxEnt in primal form} & \quad \min_{\mathbf{q} \geq 0} \quad & K(L(q||w)) \\
& \quad \text{subject to} \quad & \mathbf{q}(x) \log \frac{q(x)}{u(x)} \\
& & \mathbf{q}(x) \geq \mathbf{1} \mathbf{x} \\
& & \mathbf{q}(x) = \mathbf{1} \\
& & \mathbf{q}(x) \mathbf{T}(x) = \mathbf{0} \\
& & \mathbf{q}(x) = 1 \\
\end{align*}
\]

 absentee this constraint in domain of def. of \( K(q||w) \) i.e.

\[
\ell(q, u, \mathbf{c}) = \frac{1}{x} q(x) \log \frac{q(x)}{u(x)} + \frac{1}{\mathbf{u}} \sum_{j} \left( q_j - \mathbf{c}_j \right) \log \left( q_j / \mathbf{c}_j \right) + \mathbf{c}^T \left( \mathbf{1} - \mathbf{1} q(x) \right)
\]

\[
\begin{align*}
\frac{\partial \ell}{\partial q(x)} &= \frac{1}{x} \log \frac{q(x)}{u(x)} - \sum_{j} \frac{q_j}{u_j} \mathbf{T}(x) - \mathbf{c} = 0 \\
\Rightarrow \left( \frac{\partial \ell}{\partial \mathbf{q}(x)} \right)_C = u(x) \exp \left( \mathbf{u}^T \mathbf{T}(x) + c - 1 \right)
\end{align*}
\]
\[
\Rightarrow \quad \text{exp \text{family}}
\]

dual problem:

plug back \(q(x) \) in \( \mathcal{F}(\cdot) \)

\[
g(n,c) = \mathcal{F}(q(x), n, c)
\]

\[
= \frac{\exp\left[\mathcal{F}(q(x), n, c)\right]}{\mathcal{Z}(n)} + cn + \exp\left[\mathcal{F}(q(x), n, c)\right]
\]

\[
= \mathcal{N}(x + c) - \left(\mathcal{N}(x) \exp(\mathcal{F}(x))\right) \exp(-c)
\]

\[
\max g(n,c) \quad \nabla_c = 0 \quad \Rightarrow \quad 1 - \mathcal{Z}(n) \exp(-c) = 0
\]

\[
\Rightarrow \quad \exp(-c) = \frac{1}{\mathcal{Z}(n)}
\]

plug back \(c^*\):

\[
\max_c g(n,c) = \mathcal{N}(x + c^*) - \frac{1}{\mathcal{Z}(n)}
\]

\[
c^* = -\log \mathcal{Z}(n)
\]

\[
\text{dual problem:} \quad \max_n \tilde{g}(n) \quad \tilde{g}(n) = \mathcal{N}(x) - \log \mathcal{Z}(n)
\]

link with MLE:

if \(\alpha = \frac{1}{n} \sum_{i=1}^{n} T(x^{(i)}) = \mathbb{E}_{n} T(x)\)

then

\[
g(n) = \frac{1}{n} \sum_{i=1}^{n}[\mathcal{N}(x^{(i)}) - \log \mathcal{Z}(n)]
\]

\[
= \log p(x^{(i)}|\theta) + \text{const}
\]

\[
\text{where } p(x^{(i)}|\theta) = u(x) \exp(\mathcal{N}(x^{(i)} - \log \mathcal{Z}(n))
\]

i.e., dual problem is \(\max_n g(n) = \max_n \frac{1}{n} \log p(x^{(i)}|\theta)\)

\[
\text{i.e. MLE}\]

to summarize, ML in exp family with \(T(x)\) as sufficient statistics

is equivalent to Max BVT with moment constraints on \(T(x)\)

where \(\alpha = \mathbb{E}_{n} T(x)\).
to summarize: ML in exp family with $T(x)$ as sufficient statistics

is equivalent to MaxEVT with moment constraints on $T(x)$

where $x = \mathbb{E}_p[T(x)]$

they are Lagrangian dual of each other?

MLE in exp family $\iff$ moment matching in exp family

Note:

$\nabla_x \log Z(x) = \nabla_x \frac{\mu(x)}{Z(x)}$

$= \frac{\mu(x) \mu'(x)}{Z(x)}$

$\nabla_x \log Z(x) = \mathbb{E}_{p(x|y)} \mu'(x) \mu(x) \mu'(x) \mu(x)$

$= \frac{\mu'(x) \mu(x)}{Z(x)}$

$\nabla_x \hat{g}(x) = \left[ \mathbb{E}_{p(x|y)} \mu'(x) \mu(x) \right] - \mu'(x) \mu(x)$

$= \hat{\mu}_n$ (empirical moment)

$\nabla \hat{g}(x) = 0$

$\Rightarrow \hat{\mu}(x^*) = \hat{\mu}_n$

i.e. moment matching?

→ see lecture 16 in 2017 for "KL Pythagorean theorem"

(see end of old lecture 16 2017 for "KL Pythagorean theorem" and I-projection vs. M-projection for KL + geometry)