Lecture 18 - scribbles

Friday, November 9, 2018 13:26

today: finish expon. family
sampling

(cts. exp family)

example 3: UGM?

let $p \in \mathcal{L}(G)$, G is undirected

with $\psi_c(x_c) > 0 \quad \forall c, x_c$

$$p(x) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(x_c) = \exp\left(\sum_c \ln \psi_c(x_c) - \log Z \right)$$

$$= \exp\left(\sum_{c \in \mathcal{C}} \sum_{y_c \in \mathcal{X}_c} \underbrace{\mathbb{1}\{y_c = x_c\}}_{T_{c,y_c}(x)} \underbrace{\log \psi_c(x_c)}_{\eta_{c,x_c}} - \log Z \right)$$

$$T(x) = \begin{pmatrix} \vdots \\ \mathbb{1}\{x_c = y_c\} \\ \vdots \\ \vdots \end{pmatrix} \leftarrow y_c \in \mathcal{X}_c, \; c \in \mathcal{C}$$

$$\eta(\theta) = \begin{pmatrix} \vdots \\ \log \psi_c(y_c) \\ \vdots \end{pmatrix}$$

notes: a) Mult(π) is special case with complete graph (1 big clique)

b) feature perspective: instead of using all indicators $\mathbb{1}\{x_c = y_c\}$

you could choose relevant subset for a task

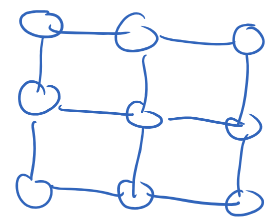
for example; if x_i is a word

feature on x_i & x_{i+1} $\mathbb{1}\{x_i \text{ is a verb } \& \ x_{i+1} \text{ is a noun}\}$

c) binary Ising model

$x_i \in \{0,1\}$ $\quad |C| \leq 2$

suppose use nodes & pairs as cliques
(edges)



$\Rightarrow$ dimension of $T(x)$ is $2|V| + 4|E|$ $\rightsquigarrow$ "overparametrized" exp. family

$$\sum_{y_c} T_{c,y_c}(x) = 1 \Rightarrow \text{not a minimal exp. family}$$

a minimal representation:

$$T(x) = \begin{pmatrix} (x_i)_{i \in V} \\ (x_i x_j)_{\{i,j\} \in E} \end{pmatrix} \rightarrow \text{dim: } |V| + |E|$$

$\mathbb{1}\{x_i = 1, x_j = 1\}$

$p_{ij} - p_i$

$x_j \backslash x_i$	0	1	
0			$1-p_j$
1	$p_{ij} - p_j$	$p_{i,j}$	p_j
	$1-p_i$	p_i	

$p_i \triangleq P\{x_i = 1\}$

properties of A:

- $A(\cdot)$ is C^∞ for $\eta \in \mathrm{int}(\Omega)$
  $A(\cdot)$ is convex (Ω is a convex set)
- $\nabla_\eta A(\eta) = \mathbb{E}_{p(x|\eta)}[T(x)] \triangleq \mu(\eta)$ "moment vector" (for $\eta \in \mathrm{int}(\Omega)$)
- $\left(\frac{\partial^2}{\partial\eta_i \partial\eta_j} A(\eta)\right)_{ij} = \mathbb{E}_{p(x|\eta)}\left[[T(x)-\mu(\eta)][T(x)-\mu(\eta)]^T\right] = \mathrm{cov}(T(X))$
  (Hessian)
  (proof as exercise?)
- $Z(\eta)$ is the mgf for $p(x;\eta)$ $\quad t \mapsto Z(\eta+t)$ for $t \in \mathbb{R}^p$
  $\mathbb{E}_X \exp(t^T T(X))$

$A(\eta) = \log \text{mgf} \rightarrow$ cumulant generating functions

$\rightsquigarrow$ why derivative of A gives cumulants

$1^{st} \rightarrow$ mean
$2^{nd} \rightarrow$ covariance
$3^{rd} \rightarrow$ 3^{rd} cumulant

14h08

Approximate inference $\rightarrow$ sampling

example: cannot do exact inference in Ising model (tree width is too big)
$\Rightarrow$ need approximations [NP-hard]

why sampling? $\quad X = (X_1, \ldots, X_p)$

a) simulation: $X^{(i)} \sim p$

b) approximate $p(x_i)$ "marginal inference"

b) approximate $p(x_j)$ "marginal inference"

$\rightarrow$ special case of <u>expectations</u>

consider: $f: \mathbb{R}^p \rightarrow \mathbb{R}^q$,

we want to approximate $\mu = \mathbb{E}_p[f(X)]$

e.g. if $f(X) = \mathbb{1}\{X_A = x_A\}$, $\mathbb{E}_p[f(X)] = p(X_A = x_A)$

<u>Monte-Carlo integration / estimation</u> $\rightarrow$ appears in physics, applied math, ML, etc...

to approximate $\mu = \mathbb{E}_p[f(X)]$

MC estimate

alg: • n samples $X^{(i)} \overset{iid}{\sim} p$

• estimate: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(X^{(i)}) = \mathbb{E}_{\hat{p}_n}[f(X)]$

<u>properties</u>: 1) unbiased

expectation over $(X^{(i)})_{i=1}^n$

this is true even if <u>$X^{(i)}$ were dependent</u>

$$\mathbb{E}[\hat{\mu}] = \frac{1}{n}\sum_i \mathbb{E} f(X^{(i)}) = \frac{\cancel{n}\mu}{\cancel{n}} \stackrel{\downarrow}{=} \mu$$

$X^{(i)} \sim p \Rightarrow X^{(i)} \overset{d}{=} X \sim p$

2) expected ℓ_2 error

$$\mathbb{E}[\|\mu - \hat{\mu}\|_2^2] = \mathbb{E}\left[\frac{1}{n^2}\sum_{i,j} \langle f(X^{(i)}) - \mu, f(X^{(j)}) - \mu \rangle\right]$$

$\|$

$tr(cov(\hat{\mu}, \hat{\mu}))$

by independence $\rightarrow$ off diagonal term are zero

$$= \frac{1}{n^2}\sum_{i=j} \underbrace{\mathbb{E}[\langle f(X^{(i)}) - \mu, f(X^{(i)}) - \mu \rangle]}_{\mathbb{E}[\|f(X^{(i)}) - \mu\|^2] \triangleq \sigma^2}$$

$= tr(cov(f(X), f(X)))$

$$\boxed{\mathbb{E}\left[\|\hat{\mu}-\mu\|^2\right] = \frac{\sigma^2}{n}}$$

$\mathbb{E}\left[\|f(x^{(i)})-\mu\|^2\right] \triangleq \sigma^2$
$= \mathrm{tr}(\mathrm{cov}(f(x), f(x)))$

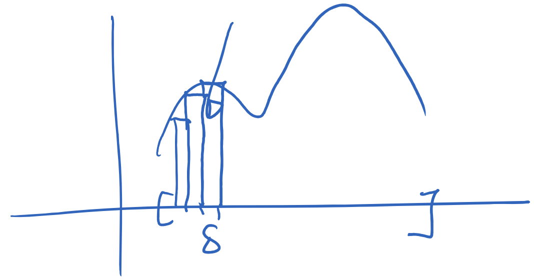
note that no dimension in rate (apart σ^2)

14h33

aside on numerical computing: 1d is "easy"

- numerical integration in 1d

error here $\leq \alpha L\delta$



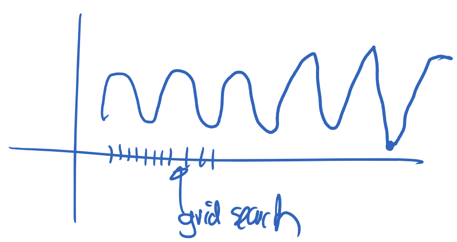
for function L-Lipschitz
$|f(x)-f(y)| \leq L\|x-y\|$

error $\leq \varepsilon \Rightarrow$ complexity $O\left(\frac{1}{\varepsilon}\right)$ of approximating integral within ε

$\rightsquigarrow$ use $\delta \approx \varepsilon$

but grid in dimension d $\quad O\left(\frac{1}{\varepsilon^d}\right) \rightarrow$ curse of dimensionality?

- global optimization in 1d



error $\propto L\delta$

Complexity (# of evaluations to find global optimum) $O\left(\frac{1}{\varepsilon}\right)$

How to sample?

1) $X \sim \text{Unif}([0,1])$ → pseudo random generator "rand"

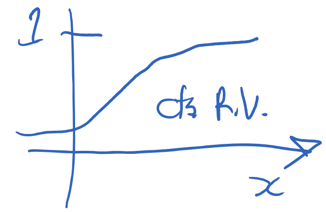
2) $X \sim \text{Bernoulli}(p)$ $\quad X = \mathbb{1}\{U \leq p\}$ where $U \sim \text{Unif}([0,1])$

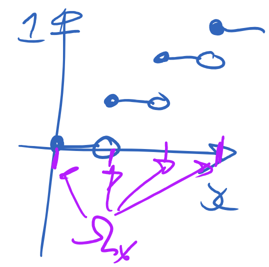
3) inverse transform sampling:

let F be target cdf (cumulative dist. fct.) of distribution p for X $\quad F(x) \triangleq P\{X \leq x\}$ $(x \in \mathbb{R})$

(first, suppose F is invertible)

let $\boxed{X \triangleq F^{-1}(U) \text{ with } U \sim \text{unif}([0,1])}$





claim that X has cdf $F(x)$

proof: $P\{X \leq y\} = P\{F^{-1}(U) \leq y\} \overset{F \text{ is invertible}}{=} P\{U \leq F(y)\} = F(y)$

[if F is not invertible, define $\boxed{X \triangleq \min\{x : F(x) \geq U\}}$] (recall that F is cts from right)

example:

want $X \sim \text{Exp}(\lambda)$ $\quad$ density $p(x) = \lambda e^{-\lambda x} \mathbb{1}_{\mathbb{R}_+}(x)$

$$F(x) = 1 - e^{-\lambda x}$$

inverse $F^{-1}(u) = -\frac{1}{\lambda} \ln(1-u)$

inverse $F^{-1}(u) = -\frac{1}{\lambda}\ln(1-u)$

Multivariate distribution?

generalize above trick using "chain rule"

$X_{1:p}$ (dim p) cdf $F(x_{1:p}) \triangleq P\{X_1 \le x_1, \dots, X_p \le x_p\}$

$$F_{X_{1:p}}(x_{1:p}) = F_{X_1}(x_1)\, F_{X_2|X_1}(x_2|x_1) \cdots F_{X_p|X_{1:p-1}}(x_p \mid x_1, \dots, x_{p-1})$$

$\hookrightarrow F_{X_2|X_1}(x_2|x_1) \triangleq P\{X_2 \le x_2 \mid X_1 \le x_1\}$

Could use $U_1, \dots, U_p \overset{iid}{\sim} \text{Unif.}$

$X_1 = F_{X_1}^{-1}(U_1)$

$\vdots$

$X_p = F^{-1}_{X_p|X_{1:p-1}}(U_p \mid X_{1:p-1})$ (curse of dimensionality?)

is very complicated fct.

[aside: "copulas" $\longrightarrow$ model for multivariate dependence with uniform marginals]

exception is multivariate Gaussian:

$N(\mu, \Sigma)$ $\Sigma = U \Lambda U^T$

(where $UU^T = I_p$) and Λ is diagonal

generate $V \sim N(0, I_p)$

$(V_p \overset{iid}{\sim} N(0,1))$

$X \triangleq U\Lambda^{1/2} V + \mu$

(Cholesky decomposition)

and Λ is diagonal

$$X \triangleq \underbrace{U\Lambda^{1/2}}_{L} V + \mu$$

(Cholesky decomposition)

$$L = U\Lambda^{1/2}$$

$$\Sigma = LL^T$$

$$\mathbb{E}X = \mu$$

$$\operatorname{cov}(X) = \Sigma$$

so $X \sim N(\mu, \Sigma)$

Box-Muller transform to sample (2d) Gaussian

$$x = r\cos\theta \qquad y = r\sin\theta$$

$\rightsquigarrow$ $r^2 \sim \mathrm{Exp}(1)$

θ is unif $([0, 2\pi])$

$$\Rightarrow \begin{pmatrix} x \\ y \end{pmatrix} \sim N(0, I)$$