

## 1 Basics

- General p-norm:  $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$
- Hoelder:  $\|uv\|_1 \leq \|u\|_p \|v\|_q$ ,  $\|u+v\|_p \leq \|u\|_p + \|v\|_p$
- Triangle:  $\|u+v\|_p \leq \|u\|_p + \|v\|_p$
- Cauchy-Schwarz:  $|\langle u, v \rangle|^2 \leq \|u\|^2 \|v\|^2$
- Cauchy-Schwarz:  $|\mathbb{E}[X, Y]|^2 \leq \mathbb{E}[X^2] \mathbb{E}[Y^2]$
- Taylor:  $f(x) \approx f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \dots$ 
  - $f(x) \approx f(a) + \frac{\partial f(x)}{\partial x} \Big|_a - \frac{1}{2}(x-a)^\top \left( \frac{\partial^2 f(x)}{\partial x \partial x^\top} \Big|_a \right) (x-a)$
  - Power series of exp.:  $\exp(x) := \sum_{k=0}^{\infty} \frac{x^k}{k!}$
- Entropy:  $H(X) \equiv H(p_X) = \mathbb{E}_X[-\log \mathbb{P}(X=x)]$ 
  - $H(X|Y) = \sum_y \mathbb{P}(Y=y) H(X|Y=y) \leq H(X)$
  - $H(X, Y) = H(X) + H(Y|X)$
  - $H(X|g(X)) \geq 0$     $H(g(X)|X) = 0$
  - $H(cX) \begin{cases} = H(X) & \text{discrete} \\ = H(X) + \log|c| > H(X) & \text{continuous} \end{cases}$
- MI:  $I(X; Y|Z) = H(X|Z) - H(X|Y, Z)$  (symmetric)
  - $I(X; Y) = D_{\text{KL}}(p(x, y) \| p(x)p(y)) \geq 0$
  - $I(X_1, \dots, X_n; Z) = \sum_{i=1}^n I(X_i; Z | X_1, \dots, X_{i-1})$
  - Markov chain:  $I(X_1; X_2, X_3, \dots) = I(X_1; X_2)$
  - $I(X, Y; Z) = I(X; Z) + I(Y; Z | X)$
- KL-divergence:  $D_{\text{KL}}(p \| q) = \sum_x p(x) \log \left( \frac{p(x)}{q(x)} \right) \geq 0$
- $1 - z \leq \exp(-z)$
- Jensen,  $f(X)$  convex:  $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$

### 1.1 Calculus

- Partial:  $\int uv' dx = uv - \int u'v dx$     $\frac{\partial}{\partial x} \frac{g}{h} = \frac{g'h - gh'}{h^2}$
- $\frac{\partial}{\partial x} (\|x-b\|_2) = \frac{x-b}{\|x-b\|_2}$     $\frac{d}{dx} |x| = \frac{x}{|x|}$
- $\frac{\partial}{\partial X} \log|X| = X^{-\top}$     $|X^{-1}| = |X|^{-1}$
- $\frac{\partial}{\partial x} (b^\top x) = \frac{\partial}{\partial x} (x^\top b) = b$
- $\frac{\partial}{\partial x} (b^\top Ax) = A^\top b$     $\frac{\partial}{\partial x} (c^\top Xb) = cb^\top$
- $\frac{\partial}{\partial X} (c^\top X^\top b) = bc^\top$     $\frac{\partial}{\partial x} (x^\top x) = 2x$
- $\frac{\partial}{\partial x} (x^\top Ax) = (A^\top + A)x \stackrel{\text{A sym.}}{=} 2Ax$
- $\frac{\partial}{\partial X} \text{Tr}(X^\top A) = A$    Trace trick:  $x^\top Ax = \dots$ 
  - $\dots \stackrel{\text{inn. prod.}}{=} \text{Tr}(x^\top Ax) \stackrel{\text{cycl. perm.}}{=} \text{Tr}(xx^\top A) = \text{Tr}(Axx^\top)$

## 2 Maximum Entropy Inference

Sample  $c \sim p(\cdot | X)$  s.t.  $H[p(\cdot | x)]$  is maximal,  
 $\mathbb{E}_{C|X}[R(C, X)] = \mu$  and  $\sum_c p(c | X) = 1$ .

$$\Rightarrow \text{Gibbs dist.: } p(c | X) = \frac{1}{Z(X)} \exp(-\beta R(c, X))$$

**Free energy:**  $F(X) := -\frac{1}{\beta} \log Z(X)$

$$\Leftrightarrow p(c | X) = \exp(-\beta[R(c, X) - F(X)])$$

$$\Rightarrow \text{entropy: } H[c | X] = \beta \underbrace{\mathbb{E}_{C|X}[R(C, X)]}_{=\mu} - \beta F(X)$$

$$\text{ME: } \max H[c | X] \Leftrightarrow \max Z(X) \Leftrightarrow \min F(X)$$

- Exp. generalisation costs:  $\mathbb{E}_{X''} \mathbb{E}_{C|X'} \mathbb{E}_{C|X''} [R(c, X'')]$
- Min. out-of-sample descr. length per deg. of freedom  

$$\min_{p(\cdot)} \mathbb{E}_{X', X''} \mathbb{E}_{C|X'} \left[ -\log \frac{p(c|X'')}{p(c)} \right] \quad p(c) = \mathbb{E}_X[p(c | X)]$$

$$\stackrel{\text{Jensen}}{\geq} \min_{p(\cdot)} \mathbb{E}_{X', X''} \left[ -\log \mathbb{E}_{C|X'} [p(c | X'')] \right] - H[c]$$

$$= \max_{p(\cdot)} \mathbb{E}_{X', X''} [e^{H[c]} \cdot \kappa(X', X'')]$$

$$\text{PA: } T^* = \arg \max_T \kappa(X', X'')$$

- PA-kernel:  $\kappa(X', X'') := \sum_c p(c | X') p(c | X'')$
- combined:  $p(c | X', X'') \propto p(c | X') p(c | X'')$

## 3 Methods for intractable Gibbs distr.

### 3.1 Markov Chains

**Mixing time of MC:**  $\|P^t(c, \cdot) - \pi(\cdot)\|_{TV} \leq \epsilon$

$t_{\text{mix}} \propto \frac{1}{\lambda_1 - \lambda_2}$  where  $1 = \lambda_1 > \lambda_2 \geq \dots$

Well behaving **Markov Chains** are

- irreducible:** can go from/to any state (n steps)
- aperiodic:** chain doesn't go back & forth forever  
 $(\forall n > n(c, c'))$  (no) path length n w/ non-zero prob.)

$$1. \wedge 2. \Rightarrow \text{unique stat. dist. } p(c') = \sum_c \pi(c | c') p(c)$$

$$1. \wedge 2. \wedge \text{stat.} \Rightarrow \lim_{t \rightarrow \infty} \mathbb{P}[X_t = c] = p(c) \wedge$$

$$\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{s=1}^t f(X_s) = \sum_c p(c) f(c)$$

$$\text{DBE } \pi(c' | c) p(c) = \pi(c | c') p(c') \Rightarrow \text{p stat.}$$

$$\text{MH: } \lambda_2 = \max \left\{ 1 - \frac{q(y, x)}{p_x}, 1 - \frac{q(x, y)}{p_y} \right\} = 1 - \alpha - \beta$$

## 3.2 Sampling and SA

**Metropolis-Hastings:** Assume  $p(c) \propto f(c)$ .

$$\pi(c' | c) := \begin{cases} q(c' | c) A(c, c') & c \neq c' \\ 1 - \sum_{c' \neq c} q(c' | c) A(c, c') & \text{otw.} \end{cases}$$

where  $q(c' | c)$ : prob. to propose the move  $c \rightarrow c'$ ,  
 and  $A(c, c') := \min \left\{ 1, \frac{q(c|c') f(c') / Z}{q(c'|c) f(c) / Z} \right\}$  prob. accept move

**Metropolis Algorithm:** Assume  $p(c) \propto f(c)$  and  
 $q(c' | c) = q(c | c')$ , i.e. symmetric.

- Define symmetric  $\{q(\cdot | c)\}_{c \in \mathcal{C}}$  s.t. graph  $G_q$  is connected and every vertex in  $G_q$  has edge to itself.
- $c_0^T \leftarrow \$$  while  $T > \epsilon$  do:
  - for  $t = 1, 2, \dots, N$ , do:
    - $\tilde{c} \leftarrow q(\cdot | c_{t-1}^T)$  // sample
    - $b \leftarrow \text{Bern} \left( \min \left\{ 1, e^{-\frac{1}{T} [R(\tilde{c}, X) - R(c_{t-1}, X)]} \right\} \right)$
    - If  $b = 1$  then  $c_t^T \leftarrow \tilde{c}$  else  $c_t^T \leftarrow c_{t-1}^T$ .
  - $c_0 \leftarrow c_N^T$
  - $T \leftarrow \text{reduce}(T)$
  - $c_0^T \leftarrow c_0$

**Temperature:** high temperature  $T \rightarrow$  closer to uniform i.e. worse ability to discriminate between good and bad models  $\rightarrow$  more likely to accept moves i.e. exploration, not stuck in bad local minima  
 reduce temperature to find better local minima and get stuck there

### 3.3 Laplace's Method (Least angle clust.)

- Square the cost:  $e^{-\frac{1}{T} R(c, X)} = \text{const} \cdot e^{g(c)^\top g(c)}$
- Complete the square:  $\int e^{-\frac{1}{T} (y - g(c))^2} dy = (\pi T)^{d/2}$   
 $\Rightarrow e^{g(c)^\top g(c)} = (\pi T)^{-d/2} \int \exp^{-y^\top y + 2y^\top g(c)} dy$
- Rewrite normalisation constant:  
 $Z = \sum_c e^{-\frac{1}{T} R(c, X)} = \dots = \text{const} \int e^{-\frac{1}{T} f(y)} dy$
- Apply Laplace's method:  
 If  $f$  has unique min.  $y_0$  and Hessian  $H := \frac{\partial^2 f}{\partial y^2} \Big|_{y_0}$   

$$\int e^{-\frac{1}{T} f(y)} dy \stackrel{(T \rightarrow 0)}{\approx} e^{-\frac{1}{T} f(y_0)} \left| \frac{H}{2\pi T} \right|^{-1/2}$$

### 3.4 Mean-field Approximation

**Idea:** Approximate  $p_\beta$  (Gibbs) with a “simple”, factorisable distribution  $p = p_1 \cdots p_N$ .

**Approach:** Minimise  $D_{KL}(p \parallel p_\beta)$

$\iff$  Minimise **Gibbs free energy**:

$$G(p) = \frac{1}{\beta} D_{KL}(p \parallel p_\beta) + F(\beta) = \mathbb{E}_{c \sim p}[R(c)] - \frac{1}{\beta} H[p]$$

Note:  $H[p] = \sum_{i=1}^N H[p_i]$  and  $F(\beta) \leq G(p)$

**Ising model:**  $E(\sigma | h) = -\frac{\beta}{2} \sum_i \frac{h_i}{|N_i|} \sum_{j \in N_i} h_j - \lambda \sum_i h_i \sigma_i$

$$E(\sigma | h) = -\sum_i^p h_i \sigma_i - \lambda \sum_i^{p-r} \sigma_i \sigma_{i+r}$$

$$E(\sigma | h) = -\sum_i h_i \sigma_i - \lambda \sum_{i,j} J_{ij} \sigma_i \sigma_j$$

where  $J_{ij}$ : interaction between particles,

$h_i$ : noisy image,  $\sigma_i$ : denoised image

**Problem:**  $\frac{\partial G(p)}{\partial p_{u\alpha}} = 0$  s.t.  $\sum_{v \leq K} p_{iv} = 1 \forall i$

**Solution:** with the *mean field*  $h_u = [\cdots h_{u\alpha} \cdots]^\top$

$$h_{u\alpha} := \frac{\partial \mathbb{E}[R(c)]}{\partial p_{u\alpha}} = \mathbb{E}_{c \sim p_{|u \rightarrow \ell}}[R(c)] \leftarrow \begin{matrix} \text{object } u \text{ chooses} \\ \text{class } \alpha \end{matrix}$$

**E:**  $p_{u\alpha} = \frac{e^{-\beta h_{u\alpha}}}{\sum_{v \leq K} e^{-\beta h_{uv}}}$  1. Pick random

$i$  2.  $h_i^{\text{new}} \leftarrow p_j^{\text{old}}$  3.  $p_i^{\text{new}} \leftarrow h_i^{\text{new}}$

**M:**  $h_{u\alpha} = \sum_c \prod_{i=1, i \neq u}^N p_i(c(i)) \mathbb{I}_{\{c(u)=\alpha\}} R(c, x)$ ,

$$-H[P] = \sum_{i \leq N} \sum_{c(i) \in \{1, \dots, K\}} \prod_{j \in N \setminus \{i\}} p_j(c(j)) \sum_{c(i) \in \{1, \dots, K\}} p_i(c(i)) \log p_i(c(i))$$

$$= \sum_{i=1}^N \sum_{v=1}^K p_i(v) \log p_i(v)$$

#### 3.4.1 Smooth k-means scr.20 (p. 39)

$$R(c | X) = \sum_i \|x_i - y_{c(i)}\|^2 + \frac{\lambda}{2} \sum_i \sum_{j \in N(i)} \mathbb{I}_{\{c_i \neq c_j\}}$$

where the second term measures **#violations** of these neighbourhood constraints.

$$\implies h_{i\ell} = \|x_i - y_\ell\|^2 + \lambda \sum_{j \in N(i)} p_{j\ell} + \text{const}_i$$

### 4 Deterministic Annealing (Z is tractable)

**Lemma:** func's  $\times$  domain  $\rightarrow$  domain  $\times$  co-dom.

$$\mathcal{O}(K^N) \rightarrow \sum_c \prod_i \epsilon_{i,c(i)} = \prod_i \sum_k \epsilon_{ik} \leftarrow \mathcal{O}(NK)$$

$$p(c | \theta, X) = \prod_{i \leq N} p_i(c(i) | \theta, X)$$

$$\text{where } p_i(k | \theta, X) \propto \exp(-\frac{1}{T} \|x_i - \theta_k\|^2)$$

$$\theta_k^* = \arg \max_{\theta_k} \left\{ \frac{\mathbb{E}[R]}{T} + \sum_{i \leq n} \log \sum_{v \leq K} \exp(-\frac{1}{T} \|x_i - \theta_v\|^2) \right\}$$

$\implies$  **Maximize Entropy**

$$\implies \frac{\partial \log Z}{\partial \theta_k} = \frac{\partial \sum_{i \leq n} \log \sum_{v \leq K} \exp(-\|x_i - \theta_v\|^2)}{\partial \theta_k} \triangleq 0 \implies$$

$$\theta_k^* = \frac{\sum_i p_i(k | \theta^*, X) \cdot x_i}{\sum_i p_i(k | \theta^*, X)}$$

do

$$\text{E-step: } p_i(k | \theta^{\text{old}}, X) = \frac{\exp(-\frac{1}{T} \|x_i - \theta_k\|^2)}{\sum_{j \leq K} \exp(-\frac{1}{T} \|x_i - \theta_j\|^2)}$$

$$\text{M-step: } \theta_k \leftarrow \frac{\sum_{i \leq n} p_{ik} x_i}{\sum_{i \leq n} p_{ik}}$$

$$\theta^{\text{old}} \leftarrow \theta$$

until convergence of  $\theta$

$$\theta_k \leftarrow \theta_k + \epsilon \quad (\text{noise s.t. centroids can separate})$$

**Phase transitions:** For  $T \rightarrow \infty$ :  $\theta_k^* = \bar{X} \quad \forall k \leq K$

Once  $T = 2\lambda_{\max}$ , more centroids appear, where

$\lambda_{\max} = \text{max. eigenvalue of } \frac{1}{N} X^\top X$ . ( $x_i$ 's row-wise)

**DA vs MAP:**

1. MAP can get stuck in local maximum
2. MAP not robust against noisy data
3. DA guaranteed to obtain global optimum if annealing is slow enough and ergodicity is given
4. In DA  $T > 0$  gives entropic regularisation

**Limiting Behaviour:**

$$\begin{aligned} \bullet \lim_{T \rightarrow \infty} : & \quad P(c(i) = k) \rightarrow \frac{1}{K}; \theta_k^* \rightarrow \frac{1}{N} \sum_i x_i \\ \bullet \lim_{T \rightarrow 0} : & \quad P(c(i) = k) \rightarrow \begin{cases} 1 & \text{if } k = \arg \min_j \|x_i - \theta_j\| \\ 0 & \text{otw.} \end{cases}; \theta_k^* \rightarrow \frac{\sum_{i \in X_k} x_i}{|X_k|} \end{aligned}$$

### 5 Histogram Clustering

**Least Angle Clust. (LAC):** [Idea]

Similarity  $S(x_i, x_j) = w_{ij} \cos(\phi_{ij}) = w_{ij} e_i \cdot e_j$  with unit vectors  $e_i := x_i / \|x_i\|$ , e.g. choice  $w_{ij} = \|x_i\| \cdot \|x_j\|$ .

**Dyadic data:**  $\mathcal{Z} = \{(x_{i(r)}, y_{j(r)}); 1 \leq r \leq \ell\}$

• prototype / “centroid”:  $q(y_j | \alpha)$

• joint dist.:  $\hat{p}(x_i, y_j) = \frac{1}{\ell} \sum_{r \leq \ell} \Delta_{x_i, x_{i(r)}} \Delta_{y_j, y_{j(r)}}$

**empirical dist.:**  $\hat{p}(y_j | x_i) = \frac{\hat{p}(x_i, y_j)}{\hat{p}(x_i)} = \frac{n(x, y)}{n(x)} \leftarrow \text{scr. (5.10)}$

**Likelihood:**  $P(\mathcal{Z} | c, q) = \prod_{r \leq \ell} p(x_{i(r)}, y_{j(r)} | c, q)$   
 $\leftarrow \text{scr. (5.11)}$   
 $= \text{scr. (5.12)} = \prod_i \prod_j [q(y_j | c(i)) \cdot p(c(i)) \cdot p(x_i)]^{\ell \hat{p}(x_i, y_i)}$

Assume  $p(\alpha) = 1/k$  and  $\hat{p}(x_i) = 1/n$

**Cost:**  $R^{\text{hc}}(c, q, \mathcal{Z}) = \frac{\ell}{n} \sum_{i \leq n} D_{KL}[\hat{p}(\cdot | x_i) \parallel q(\cdot | c(i))]$   
 $nll = -\sum_{i \leq n} \sum_{j \leq m} \ell \hat{p}(x_i, y_j) \log(q(y_j | c(i)) p(c(i)) p(x_i)) +$

$$\sum_{i \leq n} \sum_{j \leq m} \ell \hat{p}(x_i, y_j) \log \hat{p}(y_j | x_i) =$$

$$\ell \sum_{i \leq n} \sum_{j \leq m} \hat{p}(x_i, y_j) \log \frac{\hat{p}(y_j | x_i)}{q(y_j | c(i))} + K$$

Solving the **Gibbs dist.**  $p(c | q, \hat{p}) = \prod_{i \leq n} P_{i,c(i)}$

via Lagrange yields  $q^*(y_j | \alpha) = \frac{\sum_{i \leq n} P_{i\alpha} \cdot \hat{p}(y_j | x_i)}{\sum_{i \leq n} P_{i\alpha}}$  Lemma 2 ch.3 p.36

where  $\frac{\partial H}{\partial \theta} = -\frac{1}{T} \mathbb{E}_{C(\cdot | \theta, X)} \left[ \frac{\partial R(C, \theta, X)}{\partial \theta} \right]$

**Generative Model:**

1. pick random object  $x_i \in \mathcal{X}$  according to  $p(x_i)$
2. select its cluster membership  $c(i)$  of  $x_i$
3. select a feature value  $y_j$  according to  $q(y_j | c(i))$

### 5.1 Information Bottleneck Method

**Rate dist. theory:**  $R(D) = \min_{p(\hat{X}|X): \mathbb{E}_{\hat{X}}[d(X, \hat{X})] < D} I(X, \hat{X})$

$$\frac{\partial}{\partial p(\hat{x}|x)} (I(x, \hat{x}) + \beta \sum_x \sum_{\hat{x}} p(\hat{x} | x) p(x) d(x, \hat{x}) + \sum_x \lambda(x) (\sum_{\hat{x}} p(\hat{x} | x) - 1))$$

$$\implies p(\hat{x} | x) = \frac{p(\hat{x})}{Z(x, \beta)} \exp(-\beta d(x, \hat{x}))$$

Find efficient code  $X \mapsto \hat{X}$  (codebook vector) and preserve relevant info. about context  $Y$ .

**Criterion:**  $R^{\text{IB}}(q(\hat{x} | x)) = I(X; \hat{X}) - \beta I(\hat{X}; Y)$

**Markov chain:**  $\hat{X} \xrightarrow{q(\hat{x}|x)} X \xrightarrow{p(y|x)} Y$

**Generation process:** w/ distortion  $d(x, \hat{x}) = D_{KL}[\cdot]$

$$\begin{cases} q_t(\hat{x}|x) &= \frac{q_t(\hat{x})}{Z_t(x, \beta)} \cdot \exp(-\beta D_{KL}[p(y|x) \parallel p_t(y|\hat{x})]) \\ q_{t+1}(\hat{x}) &= \sum_x p(x) \cdot q_t(\hat{x} | x) \\ p_{t+1}(y|\hat{x}) &= \sum_x p(y | x) \cdot p(x) \cdot q_t(\hat{x} | x) / q_t(\hat{x}) \end{cases}$$





$$p(X|M) = \int_{\Theta_M} \exp(\log p(X|M, \theta)) \cdot p(\theta|M) d\theta$$

Assume flat prior  $p(\theta|M) \approx \text{const}$  and

expand log-likelihood by ML estimator  $\hat{\theta}$ :

$$\bar{\ell}(\theta) = \frac{\ell(\theta)}{n} = \frac{1}{n} \log p(X|M, \theta) \stackrel{\text{i.i.d.}}{=} \frac{1}{n} \sum_i \ell(\theta, X_i) \stackrel{\text{Taylor}}{\approx} \dots$$

$$\implies p(X|M) = \text{const}_2 \cdot \exp\left(\ell(\hat{\theta}) - \frac{k'}{2} \log n\right)$$

where  $k'$ : dimension of (trainable) parameters

## 8 Model Validation

### 8.1 Stability-based Validation

**Stability:** Solutions on two data sets drawn from the same source should be similar.

### 8.2 Information-theoretic Validation

#### 8.2.1 Shannon's Channel Coding Thm.

- **Channel:**  $(\mathcal{S}, \{p(\cdot|s)\}_{s \in \mathcal{S}})$ ,  $\mathcal{S}$ : alphabet
  - $\epsilon$ -noisy binary channel:  $p(\hat{s}|s) = \begin{cases} 1-\epsilon & \text{if } \hat{s}=s \\ \epsilon & \text{if } \hat{s} \neq s \end{cases}$
- **Capacity:**  $\text{cap} = \max_p I(S; \hat{S}) \rightsquigarrow p_S(s)$
- **$(M, n)$ -code:** is a pair  $(\text{Enc}, \text{Dec})$   $\leftarrow$  scr. p.87
  - where  $M$ : #messages,  $n$ : code-length
  - **Rate:**  $r = \frac{\log_2 M}{n} \Leftrightarrow M = \lfloor 2^{nr} \rfloor$
  - **Commu. err.:**  $p_{\text{err}} := \max_{i \leq M} \mathbb{P}(\text{Dec}(\widehat{\text{Enc}(i)}) \neq i)$

Goal / **Best code:**  $\lim_{n \rightarrow \infty} \frac{\log M}{n} \text{ s.t. } \lim_{n \rightarrow \infty} p_{\text{err}} \rightarrow 0$

#### Asymptotic equipartition property (AEP):

- $A_\epsilon^{(n)}$ : Typical set of sequences  $(s_1, \dots, s_n) \in \mathcal{S}^n$ 

$$\left| -\frac{1}{n} \log p_{\mathcal{S}^n}(s^n) - H[S] \right| < \epsilon \quad \leftarrow \text{scr. p.89}$$
- $\mathbb{P}\left((S^n, \hat{S}^n) \in A_\epsilon^{(n)}\right) \xrightarrow{n \rightarrow \infty} 1 \quad \leftarrow \text{scr. p.90}$
- $p_{\text{err}} \leq 2^{-n(\text{cap} - 3\epsilon - r)} \xrightarrow{n \rightarrow \infty} 0 \text{ if } r < \text{cap}$

#### 8.2.2 Algorithm Validation

**Assumptions:**

- Exponential solution space, i.e.  $\log|\mathcal{C}| = \mathcal{O}(n)$
- $\mathcal{A}$ 's output is probabilistic, i.e.  $p(\cdot|X')$

**Ideal variant:**

**Messages:**  $\mathcal{M} = \{X'_1, \dots, X'_m\}$

**Code:**  $X'_i \xrightarrow{\text{Enc}_{\mathcal{A}}} p(\cdot|X'_i) \xrightarrow{\mathcal{C}_{\mathcal{A}}} p(\cdot|X''_i) \xrightarrow{\text{Dec}_{\mathcal{A}}} \hat{X}$

**Empirical variant:**

**Messages:**  $\mathcal{M} = \{\tau_1, \dots, \tau_m\}$  drawn u.a.r. from  $\mathbb{T}$

- Require  $\sum_{\tau} p(c|\tau \circ X') \approx \frac{|\mathbb{T}|}{|\mathcal{C}|} \pm \rho \quad \leftarrow \text{scr. p.95}$

**Code:**  $\tau_i \xrightarrow{\text{Enc}} p(\cdot|\tau_i \circ X') \xrightarrow{\mathcal{C}_{\mathcal{A}}} p(\cdot|\tau_i \circ X'') \xrightarrow{\text{Dec}} \hat{\tau}$

- $\text{Enc}_{\mathcal{A}}$ : encodes  $\tau_i \in \mathcal{M}$  as  $p(\cdot|\tau_i \circ X')$
- $\text{Dec}_{\mathcal{A}}$ : selects  $\hat{\tau} = \arg \max_{\tau} \kappa(\tau_i \circ X'', \tau \circ X')$

whereby  $\kappa(X'', X') := \sum_c p(c|X'')p(c|X')$

#### Asymptotic Equipartition Property (AEP):

AEP fulfilled if  $\log \kappa(X', X'') \xrightarrow{n \rightarrow \infty} \mathcal{E}$

whereby  $\mathcal{E} := \mathbb{E}_{X', X''}[\log \kappa(X', X'')]$

- $A_\epsilon^{(n)}$ : set of  $(\epsilon, n)$ -typical pairs  $X', X''$ 

$$|\log \kappa(X', X'') - \mathcal{E}| < \epsilon$$
- $p_{\text{err}} \leq P_{(n)}$  c.f. scr. (6.19)  $\xrightarrow{n \rightarrow \infty} 0$  if  $\frac{\log m}{\log |\mathcal{C}|} < I$ 

where  $I := \frac{1}{\log |\mathcal{C}|} \mathbb{E}_{X', X''}[\log(|\mathcal{C}| \kappa(X', X''))]$

### 8.3 Applications of PA

**PA:** quantifies the amount of information that algorithms extract from phenomena.  $\rightarrow$  quantified by **capacity** (max. # distinguishable messages that can be communicated)

**Temperature:**  $T^* = \arg \max_T \kappa(X', X'')$

**Cost functions:** Given  $R_1(\cdot, \cdot), \dots, R_s(\cdot, \cdot)$ 

$$\max_{\ell \leq s} \kappa_{\ell}(X', X'') = \max_{\ell \leq s} \frac{1}{Z_{X'} Z_{X''}} \sum_c e^{-\frac{1}{T} R_{\ell}(c, X')} e^{-\frac{1}{T} R_{\ell}(c, X'')}$$

**Algorithms:** Many MST (min. spanning tree) algo's are **contractive** ( $\rightarrow$  sequence of candidate sol's).

**Approximation Set Coding (ASC):**

$$p^{\text{ASC}}(c|X') = \begin{cases} 1/|G_{\gamma}(X')| & \text{if } c \in G_{\gamma}(X') \\ 0 & \text{otw.} \end{cases}$$

$$G_{\gamma}(X') := \left\{ c \in \mathcal{C} : R(c, X') - \min_{c \in \mathcal{C}} R(c, X') \leq \gamma \right\}$$

1. Run  $\mathcal{A}$  to compute  $G_t^A(X')$  and  $G_t^A(X'')$ , for all  $t$

$$2. t^* = \arg \max_t \kappa(X', X'') = \arg \max_t \frac{|G_t^A(X') \cap G_t^A(X'')|}{|G_t^A(X')| \cdot |G_t^A(X'')|}$$

$$3. c^* \xleftarrow{\$ \text{ sample}} \text{Unif}(G_{t^*}^A(X') \cap G_{t^*}^A(X''))$$

## 9 Appendix

### 9.1 Tips and Tricks

#### Complete the square:

If  $p(x) \propto \exp(-\frac{1}{2} x^T A x + x^T b)$ ,

then  $p(x) = \mathcal{N}(x | A^{-1}b, A^{-1})$

#### Constrained optimisation:

**primal:**  $\min_x f(x) \text{ s.t. } g_i(x) = 0; h_j(x) \leq 0$

**Lagrangian:** with each  $\alpha_j \geq 0$

$$\mathcal{L}(x, \lambda, \alpha) = f(x) + \sum_i \lambda_i g_i(x) + \sum_j \alpha_j h_j(x)$$

Solve:  $\frac{\partial \mathcal{L}}{\partial x} = 0; g_i(x) = 0; \alpha_j \geq 0; h_j(x) \leq 0$

If **Slater's cond.** holds,  $\exists x : g_i(x) = 0, h_j(x) < 0$ , then we can solve the **dual** instead:

$$\max_{\lambda, \alpha} \{ \min_x \mathcal{L}(x, \lambda, \alpha) \} \text{ s.t. } \alpha_j \geq 0$$

Solve:  $\frac{\partial \mathcal{L}}{\partial x} = 0; \frac{\partial \mathcal{L}}{\partial \lambda} = 0; \alpha_j h_j(x) = 0; \alpha_j \geq 0$

**Euler-Lagrange:** Find extrema of functional  $\mathcal{F}[f] =$

$$\int G(x, f(x), f'(x)) dx, \text{ thus } \frac{\partial \mathcal{F}}{\partial f} \stackrel{!}{=} 0.$$

If  $G$  is twice diff'able, then

$$\frac{\partial \mathcal{F}}{\partial f} = \frac{\partial G}{\partial f(x)} - \frac{d}{dx} \left( \frac{\partial G}{\partial f'(x)} \right) \stackrel{(*)}{=} \frac{\partial G}{\partial f(x)}.$$

(\*): when  $G$  does not depend on  $f'$ .

### 9.2 Approximations

**Laplace Approximation:**  $\frac{df}{dx} \Big|_{x_0} = 0$

$$\implies \int_{\mathbb{R}} e^{Cf(x)} dx \approx \sqrt{2\pi C} \cdot |f''(x_0)| \cdot e^{Cf(x_0)}$$

#### Hyperbolic Functions:

- $\sinh(x) = \frac{e^x - e^{-x}}{2}, \quad \frac{d}{dx} \sinh(x) = \cosh(x)$
- $\cosh(x) = \frac{e^x + e^{-x}}{2}, \quad \frac{d}{dx} \cosh(x) = \sinh(x)$
- $\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \quad \cosh^2(x) + \sinh^2(x) = 1$
- $\frac{d}{dx} \tanh(x) = 1 - \tanh^2(x) = \frac{1}{\cosh^2(x)} = \text{sech}^2(x)$