**Parameter Estimation**

*Bayesian Parameter Estimation*

Given an observation $y$ drawn from $p_\theta(y)$, how do we estimate the parameter $\theta$?
We assume prior statistical knowledge of $\theta$, namely, $w(\theta)$.

1. Terminology/Formulae:

   (a) Cost function

   $$C[\hat{\theta},\theta] \;=\; \text{cost of choosing } \hat{\theta} \text{ when true parameter is } \theta.$$

   (b) Conditional risk

   $$R_\theta(\hat{\theta}) \;=\; \mathbf{E}\{C[\hat{\theta}(Y),\theta]\}$$

   (c) Bayes risk

   $$\begin{aligned} r(\hat{\theta}) &= \mathbf{E}\{R_\Theta(\hat{\theta})\} \\ &= \mathbf{E}\{C[\hat{\theta}(Y),\Theta]\} \\ &= \mathbf{E}\big\{\underbrace{\mathbf{E}\{C[\hat{\theta}(Y),\Theta]\,|Y=y\}}_{\text{posterior cost}}\big\} \end{aligned}$$

   $\rightsquigarrow$ Minimizing the Bayes risk is equivalent to minimizing the posterior cost.

   (d) Bayes estimator

   $$\hat{\theta}_B(y) \;=\; \arg\min_{\hat{\theta}(y)\in\Lambda} \mathbf{E}\{C[\hat{\theta}(Y),\Theta]|Y=y\} \;=\; \arg\min_{\hat{\theta}(y)\in\Lambda} \int_\Lambda C[\hat{\theta}(y),\theta]w(\theta|y)d\theta$$

   (e) Conditional pdf

   $$w(\theta|y) \;=\; \frac{p_\theta(y)w(\theta)}{\int_\Lambda p_\theta(y)w(\theta)d\theta}$$

   (f) Minimum mean squared error (MMSE) estimator

   $$\begin{aligned} C[\hat{\theta},\theta] &= (\hat{\theta}-\theta)^2 \\ \hat{\theta}_{\text{MMSE}}(y) &= \mathbf{E}\{\Theta|Y=y\} \;=\; \text{conditional mean} \\ r\big(\hat{\theta}_{\text{MMSE}}\big) &= \mathbf{E}\{\mathbf{Var}(\Theta|Y)\} \;=\; \text{MMSE} \end{aligned}$$

   (g) Minimum mean absolute error (MMAE) estimator

   $$\begin{aligned} C[\hat{\theta},\theta] &= |\hat{\theta}-\theta| \\ \hat{\theta}_{\text{MMAE}}(y) &= \big\{\hat{\theta} : \Pr\{\Theta<\hat{\theta}|Y=y\} \;=\; \Pr\{\Theta>\hat{\theta}|Y=y\}\big\} \;=\; \text{conditional median} \\ &= \left\{\hat{\theta} : \int_{\hat{\theta}}^{\infty} w(\theta|y)d\theta = \frac{1}{2}\right\} \text{ when } w(\theta|y) \text{ is continuous.} \end{aligned}$$

   (h) Maximum a posteriori estimator

   $$\begin{aligned} C[\hat{\theta},\theta] &= \begin{cases} 0 & |\hat{\theta}-\theta| \le \Delta \\ 1 & |\hat{\theta}-\theta| > \Delta \end{cases} \quad \text{as } \Delta\to 0. \\ \hat{\theta}_{\text{MAP}}(y) &= \arg\max_{\hat{\theta}\in\Lambda} w(\hat{\theta}|y) \;=\; \text{conditional mode} \end{aligned}$$

   $$\text{MAP equations :} \qquad \frac{\partial}{\partial\theta}\log p_\theta(y)\Big|_{\theta=\hat{\theta}_{\text{MAP}}} + \frac{\partial}{\partial\theta}\log w(\theta)\Big|_{\theta=\hat{\theta}_{\text{MAP}}} = 0$$

(i) Vector parameters ($\underline{\theta} \in \mathbb{R}^m, \underline{y} \in \mathbb{R}^n$)

i. *Decomposable cost functions:*

- Defined as costs which can be written $C[\underline{\hat{\theta}}, \underline{\theta}] = \sum_{i=1}^{m} C_i[\hat{\theta}_i, \theta_i]$.
- Popular examples include $p$-norm: $\|\underline{\hat{\theta}} - \underline{\theta}\|_p^p$.
- MMSE: $C[\underline{\hat{\theta}}, \underline{\theta}] = \|\underline{\hat{\theta}} - \underline{\theta}\|_2^2$, and $\hat{\theta}_i = \hat{\theta}_{\mathsf{MMSE},i}$.
- MMAE: $C[\underline{\hat{\theta}}, \underline{\theta}] = \|\underline{\hat{\theta}} - \underline{\theta}\|_1^1$, and $\hat{\theta}_i = \hat{\theta}_{\mathsf{MMSE},i}$.
- Decoupled MAP: Does not preserve the character of scalar MAP.

ii. *Nondecomposable cost functions:*

- Vector MAP:
  - $C[\underline{\hat{\theta}}, \underline{\theta}] = \begin{cases} 0 & \|\underline{\hat{\theta}} - \underline{\theta}\|_\infty = \max_{1 \le i \le m} |\hat{\theta}_i - \theta_i| \le \Delta \\ 1 & \text{else} \end{cases}$
  - Analogy: vector MAP $\leftrightarrow$ frame error, while decoupled MAP $\leftrightarrow$ bit error.
- Weighted-$\ell_2$:
  - $C[\underline{\hat{\theta}}, \underline{\theta}] = \|\underline{\hat{\theta}} - \underline{\theta}\|_A^2 = (\underline{\hat{\theta}} - \underline{\theta})^t A (\underline{\hat{\theta}} - \underline{\theta})$ for positive definite symmetric $A$
  - $\underline{\hat{\theta}} = \mathbf{E}\{\underline{\theta} | Y = y\}$
  - $r(\underline{\hat{\theta}}) = \mathrm{tr}\{A\, \mathbf{E}\{\mathbf{Cov}(\underline{\theta}|Y)\}\}$

(j) Vector Gaussian statistics

$$\begin{aligned} \begin{bmatrix} \underline{Y} \\ \underline{\Theta} \end{bmatrix} &\sim \mathcal{N}\left( \begin{bmatrix} \mu_Y \\ \mu_\Theta \end{bmatrix}, \begin{bmatrix} \Sigma_Y & \Sigma_{\Theta Y}^t \\ \Sigma_{\Theta Y} & \Sigma_\Theta \end{bmatrix} \right) \\ \Rightarrow \quad \underline{\Theta}|\underline{Y} &\sim \mathcal{N}(m, \Sigma) \\ m &= \mu_\Theta + \Sigma_{\Theta Y} \Sigma_Y^{-1} (\underline{Y} - \mu_Y) \\ \Sigma &= \Sigma_\Theta - \Sigma_{\Theta Y} \Sigma_Y^{-1} \Sigma_{\Theta Y}^t \end{aligned}$$

(k) Linear observation model

$$\underline{Y} = H\underline{\Theta} + \underline{N} \qquad \text{with} \qquad \underline{\Theta} \sim \mathcal{N}(\mu_\Theta, \Sigma_\Theta) \perp \underline{N} \sim \mathcal{N}(\underline{0}, \Sigma_N)$$

$$\begin{aligned} \Rightarrow \quad \underline{\Theta}|\underline{Y} &\sim \mathcal{N}(m, \Sigma) \\ m &= \mu_\Theta + \Sigma_\Theta H^t (H\Sigma_\Theta H^t + \Sigma_N)^{-1} (\underline{Y} - H\mu_\Theta) \\ \Sigma &= \Sigma_\Theta - \Sigma_\Theta H^t (H\Sigma_\Theta H^t + \Sigma_N)^{-1} H\Sigma_\Theta \end{aligned}$$

2. Notes:

(a) Centers of the distribution

The conditional mean, conditional median and the conditional maximum are all measures of the "center" of a distribution.

If the conditional pdf $w(\theta|y)$ is *symmetric* and *unimodal* then $\hat{\theta}_{\mathsf{MMSE}}(y) = \hat{\theta}_{\mathsf{MMAE}}(y) = \hat{\theta}_{\mathsf{MAP}}(y)$.

(b) Balancing priors and observations

The Bayesian estimators weigh two types of information for making an estimate: prior knowledge of $\theta$ from $w(\theta)$ and observed knowledge of $\theta$ from $p_\theta(y)$.