L1 - Key equations (compression)

- data: $\underline{x}_1, \dots, \underline{x}_n$ $\underline{x}_i \in \mathbb{R}^d$ #1 #2
- iid Gaussian
- ML #3
- density $p(\underline{x}|\underline{\mu}, \underline{\underline{\Sigma}}) = \frac{1}{(2\pi)^{d/2}|\underline{\underline{\Sigma}}|^{1/2}} \exp\left(-\frac{1}{2}(\underline{x}-\underline{\mu})^T \underline{\underline{\Sigma}}^{-1}(\underline{x}-\underline{\mu})\right)$
- moments: $\mathbb{E}[\underline{x}] = \int_{\mathbb{R}^d} \underline{x}\, p(\underline{x}|\underline{\mu}, \underline{\underline{\Sigma}})\, d\underline{x} = \underline{\mu}$ $\quad \text{cov}[\underline{x}] = \mathbb{E}[(\underline{x} - \mathbb{E}[\underline{x}])(\underline{x} - \mathbb{E}[\underline{x}])^T]$
  $= \mathbb{E}[\underline{x}\underline{x}^T] - \mathbb{E}[\underline{x}]\mathbb{E}[\underline{x}]^T = \underline{\underline{\Sigma}}$
- probabilistic model $p(\underline{x}|\underline{\theta})$ / distri family $p(\cdot)$
- Gaussian distri family: $p(\underline{x}|\underline{\theta})$ $\underline{\theta} = \{\underline{\mu}, \underline{\underline{\Sigma}}\}$
- iid: $\underline{x}_i \overset{iid}{\sim} p(\underline{x}|\underline{\theta})$ $i = 1, \dots, N$
- Joint obs:- $p(\underline{x}_1, \dots, \underline{x}_n|\underline{\theta}) = \prod_{i=1}^N p(\underline{x}_i|\underline{\theta})$ (cont'd)
- ML: $\hat{\underline{\theta}}_{ML} = \text{argmax}\, p(\underline{x}_1, \dots, \underline{x}_n|\underline{\theta})$ via $\nabla_{\underline{\theta}} \prod_{i=1}^N p(\underline{x}_i|\underline{\theta}) = 0$
- log-trick: $\ln\left(\prod_i f_i\right) = \sum_i \ln f_i$ and $\underset{y}{\text{argmax}} \ln g(y) = \underset{y}{\text{argmax}}\, g(y)$
- ML (log-cond. like): $\underset{\underline{\theta}}{\text{argmax}} \sum_{i=1}^N \ln p(\underline{x}_i|\underline{\theta}) \Leftrightarrow \sum_{i=1}^N \nabla_{\underline{\theta}} \ln p(\underline{x}_i|\underline{\theta}) = 0$
- analytic, numerical, approx
- MVGMLE: distri family $p(\cdot)$ - Gaussians on $\mathbb{R}^d$ with unknown $\underline{\mu} \in \mathbb{R}^d$, $\underline{\underline{\Sigma}} \in \mathbb{R}^{d\times d}$, $\in S_{++}^d$
  iid $\underline{x}_i \overset{iid}{\sim} p(\underline{x}|\underline{\mu}, \underline{\underline{\Sigma}})$ $\forall i = 1, \dots, N$
- MVGMLE: $\sum_{i=1}^N \nabla_{(\underline{\mu}, \underline{\underline{\Sigma}})} \ln p(\underline{x}_i|\underline{\mu}, \underline{\underline{\Sigma}}) = 0$
- ML estimates: $\hat{\underline{\mu}}_{ML} = \frac{1}{n}\sum_{i=1}^N \underline{x}_i$ $\quad \hat{\underline{\underline{\Sigma}}}_{ML} = \frac{1}{n}\sum_{i=1}^N (\underline{x}_i - \hat{\underline{\mu}}_{ML})(\underline{x}_i - \hat{\underline{\mu}}_{ML})^T$

(*) - compression should highlight areas to revisit/tighten (use feedback adaptively)

L2 - Key equations (ln) (cont)

Regression: data: $\underline{x} \in \mathbb{R}^d$, y
goal: $f: \mathbb{R}^d \to \mathbb{R}$. $y \hat{=} f(\underline{x}; \underline{w})$ for data pair $(\underline{x}, y)$, $f(\underline{x}; w)$ - (parametric reg fn.)
uncerity: f (reg/pred) is linear in unknown $\underline{w}$

- LR model $y_i \hat{=} f(\underline{x}_i; \underline{w}) = w_0 + \sum_{j=1}^d x_{ij} w_j$
  - use $(\underline{x}_i, y_i)$ $\forall n = 1, \dots, N$ to estimate unknown $\underline{w}$: $y \hat{=} f(\underline{x}_i, \hat{w})$
- LS objective: $\underline{w}_{LS} = \underset{\underline{w}}{\text{argmin}} \sum_{i=1}^N (y_i - f(\underline{x}_i; \underline{w}))^2 = \underset{\underline{w}}{\text{argmin}}\, L$
- Geometric interp: $(\underline{x} \in \mathbb{R}^2)$ $y \in \mathbb{R}$ $\hat{\underline{w}}_{LS} = (\hat{w}_0, \hat{w}_1, \hat{w}_2)$ defines a 2-d hyperplane in $y - \underline{x}_1 - \underline{x}_2$ space
  L (loss) $= \sum_{i=1}^N \epsilon_i^2$

LR model: $y_i = w_0 + \sum_{j=1}^{d} x_{ij} w_j + \epsilon_i$; $\operatorname{argmin}_{(w_0, w_1, \ldots, w_d)} L = \sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{N} (y_i - w_0 - \sum_{j=1}^{d} x_{ij} w_j)^2$

- compact rep: $\underline{x}_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{id} \end{bmatrix}$ $\underline{X} = \begin{bmatrix} x_{11} & \cdots & x_{1d} \\ & \vdots & \\ x_{n1} & \cdots & x_{nd} \end{bmatrix} = \begin{bmatrix} -\underline{x}_1^T- \\ -\underline{x}_2^T- \\ \vdots \\ -\underline{x}_N^T- \end{bmatrix}$ $\underline{x} \in \mathbb{R}^d$ $(d < n)$

(bias w_0) $\Rightarrow$ $\underline{x}_i = \begin{bmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{id} \end{bmatrix}$ $\underline{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1d} \\ 1 & & \vdots & \\ \vdots & & & \\ 1 & x_{n1} & & x_{nd} \end{bmatrix} = \begin{bmatrix} 1 & -\underline{x}_1^T- \\ 1 & -\underline{x}_2^T- \\ \vdots & \\ 1 & -\underline{x}_N^T- \end{bmatrix}$ $\underline{w} = \begin{bmatrix} w_0 \\ \vdots \\ w_d \end{bmatrix}$ $\underline{w} \in \mathbb{R}^{d+1}$

LS repre: i) $L = \sum_{i=1}^{N} (y_i - w_0 - \sum_{j=1}^{d} x_{ij} w_j)^2 \rightarrow$ ii) $L = \sum_{i=1}^{N} (y_i - \underline{x}_i^T \underline{w})^2 \rightarrow$ iii) $L = \|\underline{y} - \underline{X}\underline{w}\|_2^2$

LS estimate of $\underline{w}$: $\hat{\underline{w}}_{LS} = \left(\sum_{i=1}^{N} \underline{x}_i \underline{x}_i^T \right)^{-1} \left(\sum_{i=1}^{N} y_i \underline{x}_i \right) \Leftrightarrow (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{y}$

$\underline{X}^T \underline{y} = \sum_{i=1}^{N} y_i \underline{x}_i$

$\underline{X}^T \underline{X} = \sum_{i=1}^{N} \underline{x}_i \underline{x}_i^T$

LS prediction: $y_{new} \hat{=} \underline{x}_{new}^T \hat{\underline{w}}_{LS}$

Full rank requirement: $(\underline{X}^T \underline{X})^{-1}$ must be full-rank (i.e. when $\underline{X} \in \mathbb{R}^{n \times (d+1)}$ has at least $(d+1)$ linearly independent rows $\rightarrow$ any point in $\mathbb{R}^{d+1}$ reached by weighted combo of $(d+1)$ rows of $\underline{X}$)

- $n < (d+1) \rightarrow$ no LS
- $(\underline{X}^T \underline{X})^{-1}$ no exist $\rightarrow$ infinite sol for "$\hat{\underline{w}}_{LS}$"
- $n \gg d$

Polynomials / non-linear bfs: linearity in param $\underline{w} \Rightarrow$ non-linear basis functions $\phi(\underline{x})$ LS invariant $\rightarrow$ pre-processing different

- see Bishop: pth order polys for (x_i, y_i) $x \in \mathbb{R}$ $y \in \mathbb{R}$

$x \in \mathbb{R}$ $y \in \mathbb{R}$ (i.e. 1d (x, y)) (x_i, y_i) $n = 1, \ldots, N$

$\underline{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & & & & \\ 1 & x_n & x_n^2 & \cdots & x_n^p \end{bmatrix}$ $\underline{w}_{LS} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{y}$ $w_{LS} \in \mathbb{R}^{(p+1)}$

$y = w_0 + w_1 x + w_2 x^2 + \ldots + w_p x^p$

- 2nd/3rd order: $y_i = w_0 + w_1 x_{i1} + w_2 x_{i2} + w_3 x_{i1}^2 + w_4 x_{i2}^2$ (order 2, dim + 1)

$y_i = w_0 + w_1 x_{i1} + w_2 x_{i2} + w_3 x_{i1}^2 + w_4 x_{i2}^2 + w_5 x_{i1}^3 + w_6 x_{i2}^3$

width $\underline{X}$ - grows as (order) $\times$ (dim + 1) (expand $\underline{X} \rightarrow$]

- Further extensions

- For $\underline{x}_i \in \mathbb{R}^{d+1}$, LSLR valid for $f(\underline{x}_i; \underline{w})$ $y_i \hat{=} f(\underline{x}_i; \underline{w}) = \sum_{s=1}^{S} g_s(\underline{x}_i) w_s$ S - no. of basis functions
  - $g_s(\underline{x}_i)$ - nonlinear f. (i.e. x_{ij}^2, $\log x_{ij}$, $\mathbb{I}(x_{ij} < a)$, $\mathbb{I}(x_{ij} < x_{ij}')$) etc.] can concatenate into $\underline{w}^T \underline{g}(\underline{x})$
- only require linearity in parameters $(w_1, \ldots, w_S)$
- construct $\underline{X}$ by putting each transformed $\underline{x}_i$ on row i
- $w_{LS} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{y}$

Geometry: - $w_{LS} = \operatorname{argmin}_{\underline{w}} \|\underline{y} - \underline{X}\underline{w}\|_2^2 = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{y}$

- w_{LS}: $\underline{y} \hat{=} \underline{X}\underline{w}$ ~~as close~~ ($\underline{X}\underline{w}$ 'close' to $\underline{y}$ as possible (Euclidean)

- $\underline{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1d} \\ \vdots & & & \\ 1 & x_{n1} & \cdots & x_{nd} \end{bmatrix} = \begin{bmatrix} 1 & - & \underline{x}_1^T & - \\ & & \vdots & \\ 1 & - & \underline{x}_N^T & - \end{bmatrix} = \begin{bmatrix} 1 & & 1 \\ \underline{x}_1 & \cdots & \underline{x}_d \\ 1 & & 1 \end{bmatrix}$ $X_{i,:} = [-\underline{x}_i^T -] \quad \forall i = 1, \ldots, N$
$X_{:,d} = \begin{bmatrix} \underline{x}_d \\ 1 \end{bmatrix} \quad \forall d = 1, \ldots, d$

- Analogous to design matrix

- $\underline{X}\underline{w} = \begin{bmatrix} 1 & 1 & & 1 \\ \underline{x}_1 & \cdots & \underline{x}_d \\ 1 & & 1 \end{bmatrix} \begin{bmatrix} w_0 \\ \vdots \\ w_d \end{bmatrix} = \sum_{j=1}^{d+1} w_j X_{:,j} = \sum_{j=1}^{d+1} w_j \underline{x}_j = w_0 \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} + w_1 \begin{bmatrix} \underline{x}_1 \\ 1 \end{bmatrix} + \ldots + w_d \begin{bmatrix} \underline{x}_d \\ 1 \end{bmatrix}$

- weight columns of $\underline{X}$ by values in $\underline{w}$ to approx $\underline{y}$
- For each column $X_{:,d} \in \mathbb{R}^N$, (d+1) columns (weighted combo) span subspace $\mathbb{R}^{d+1} \subseteq \mathbb{R}^N$
- closest point in that subspace: orthonormal projection of $\underline{y}$ onto column space of $\underline{X}$
- $\hat{\underline{y}} = \underline{X}\underline{w}_{LS} = \underline{X}(\underline{X}^T\underline{X})^{-1}\underline{X}^T\underline{y}$
- differences in visual representations ⊛ - Bishop?

L3 Key equations

- LS probabilistic: Assume: $\underline{\Sigma} = \sigma^2 \underline{I}$ $\underline{\mu} = \underline{X}\underline{w}$ $p(\underline{y}|\underline{\mu}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(\underline{y}-\underline{\mu})^T(\underline{y}-\underline{\mu})\right)$ G.D. (n-dim)

- ML for Gaussian LR: $\hat{\underline{w}}_{ML} = \underset{\underline{w}}{\operatorname{argmax}} \ln p(\underline{y}|\underline{\mu}=\underline{X}\underline{w}, \sigma^2) = \underset{\underline{w}}{\operatorname{argmax}} -\frac{1}{2\sigma^2}\|\underline{y}-\underline{X}\underline{w}\|_2^2 - \frac{n}{2}\ln(2\pi\sigma^2)$

- LS-ML equivalence: LS: $\underset{\underline{w}}{\operatorname{argmin}} \|\underline{y}-\underline{X}\underline{w}\|_2^2 \Leftrightarrow$ ML: $\underset{\underline{w}}{\operatorname{argmax}} -\frac{1}{2\sigma^2}\|\underline{y}-\underline{X}\underline{w}\|_2^2$

- LS-ML equiv, $\underline{\Sigma} = \sigma^2\underline{I}$, $\underline{\mu} = \underline{X}\underline{w}$ $\Leftrightarrow$ Gaussian iid $\epsilon_i = y_i - \underline{x}_i^T\underline{w}$
  i) $y_i = \underline{x}_i^T\underline{w} + \epsilon_i$ $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ $i = 1, \ldots, n$
  ii) $y_i \overset{ind}{\sim} N(\underline{x}_i^T\underline{w}, \sigma^2)$ $i = 1, \ldots, n$
  iii) $\underline{y} \sim N(\underline{X}\underline{w}, \sigma^2\underline{I})$

- modelling assumption: $\underline{y} \sim N(\underline{X}\underline{w}, \sigma^2\underline{I})$
- Expectation and variance of LS and ML estimators under Gaussian assumptn :-
  $\mathbb{E}[\hat{\underline{w}}_{ML}] = \underline{w}$ (unbiased) $\mathrm{Var}[\hat{\underline{w}}_{ML}] = \sigma^2(\underline{X}^T\underline{X})^{-1}$ $\mathrm{Var}[\underline{y}] = \mathbb{E}[\underline{y}\underline{y}^T] - \underline{\mu}\underline{\mu}^T = \underline{\Sigma}$ (GTT)
  $\Rightarrow \mathbb{E}[\underline{y}\underline{y}^T] = \underline{\Sigma} + \underline{\mu}\underline{\mu}^T$

- sensitivity of $\hat{\underline{w}}_{ML} \rightarrow \mathrm{Var}[\hat{\underline{w}}_{ML}] = \sigma^2(\underline{X}^T\underline{X})^{-1}$

- prediction: $\hat{\underline{y}} = \underline{X}\hat{\underline{w}}_{ML} = \underline{X}\hat{\underline{w}}_{LS}$ or $\hat{y}_i = \underline{x}_i^T\hat{\underline{w}}_{ML} = \underline{x}_i^T\hat{\underline{w}}_{LS}$ (i) (ii)

- Regularisation (general): $\underline{w}_{OPT} = \underset{\underline{w}}{\operatorname{argmin}} \|\underline{y}-\underline{X}\underline{w}\|_2^2 + \lambda g(\underline{w})$ $\lambda > 0$ - reg. param
  $g(\underline{w})$ - penalty imposing prop on $\underline{w}$

- Ridge reg :- $\hat{\underline{w}}_{RR} = \underset{\underline{w}}{\operatorname{argmin}} \|\underline{y}-\underline{X}\underline{w}\|_2^2 + \lambda\|\underline{w}\|_2^2$ i.e. $g(\underline{w}) = \|\underline{w}\|_2^2$ (penalise large $\underline{w}$)

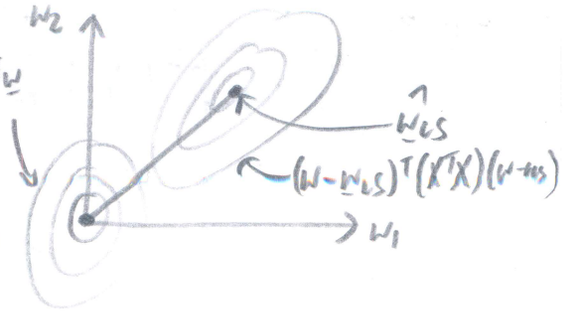
- Tradeoff: $\lambda \rightarrow 0$: $\hat{\underline{w}}_{RR} \rightarrow \hat{\underline{w}}_{LS}$
  $\lambda \rightarrow \infty$: $\hat{\underline{w}}_{RR} \rightarrow \underline{0}$

- Ridge reg estimate: $L = \|\underline{y}-\underline{X}\underline{w}\|_2^2 + \lambda\|\underline{w}\|_2^2 = (\underline{y}-\underline{X}\underline{w})^T(\underline{y}-\underline{X}\underline{w}) + \lambda\underline{w}^T\underline{w}$

$\nabla_{\underline{w}} L = \underline{0} \Rightarrow \hat{\underline{w}}_{RR} = (\lambda\underline{I} + \underline{X}^T\underline{X})^{-1}\underline{X}^T\underline{y}$

- **Geometry of ridge reg:** $\|y - Xw\|_2^2 + \lambda\|w\|_2^2$

$= (w - w_{LS})^T (X^T X)(w - w_{LS}) + \lambda w^T w + \text{const wrt } w$



- **Ridge → preprocessing:** $y \leftarrow y - \frac{1}{n}\sum_{i=1}^{N} y_i$ (subtract mean from y)

$x_{ij} \leftarrow \frac{(x_{ij} - \bar{x}_{\cdot j})}{\hat{\sigma}_j}$ $\quad \hat{\sigma}_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{ij} - \bar{x}_{\cdot j})^2}$ (standardise dimensions of x_i before X constn)

- **LS vs ridge estimates:** (parameter estimates)

$\hat{w}_{LS} = (X^T X)^{-1} X^T y$

$\hat{w}_{RR} = (\lambda I + X^T X)^{-1} X^T y$

- **SVD analysis:** $X \in \mathbb{R}^{n \times d} \Rightarrow X = USV^T$ $\quad U \in \mathbb{R}^{n \times d}$, $S \in \mathbb{R}^{d \times d}$, $V^T \in \mathbb{R}^{d \times d}$

$U^T U = I$ (col. ortho.) $\quad S_{ii} \geq 0$, $S_{ij} = 0$ $\quad V^T V = VV^T = I$ (orthonormal) $\hookrightarrow V^T = V^{-1}$

- $X^T X = VS^2 V^T$
- $(X^T X)^{-1} = (VS^2 V^T)^{-1} = VS^{-2}V^T$ (if X is full-rank $S_{ii} \neq 0 \ \forall i$)

(I)

- **SVD analysis:** (ML estimator variance) $\quad \text{var}[\hat{w}_{ML}] = \sigma^2 (X^T X)^{-1} = \sigma^2 VS^{-2}V^T \quad S_{ii}^{-2} = \frac{1}{S_{ii}^2}$
  - $S_{ii} \approx 0 \Rightarrow S_{ii}^{-2} \to \infty \Rightarrow \text{var}[\hat{w}_{ML}] \to \infty$ (columns $X_{:,j}$ highly correlated)
  - (LS prediction) $y_{new} \approx x_{new}^T \hat{w}_{LS} = x_{new}^T (X^T X)^{-1} X^T y = x_{new}^T VS^{-2}V^T VS^T U^T y = x_{new}^T VS^{-1}U^T y$
    - $S_{ii}^{-1} = \frac{1}{S_{ii}}$ (small S_{ii} (large)); unstable predict. $\quad \hat{w}_{LS} = VS^{-1}U^T y$

- **SVD analysis (II):-** (conditions under which SVD of $\hat{w}_{LS}$ and $\hat{w}_{RR}$ are the same)

$\hat{w}_{RR} = (\lambda(X^T X)^{-1} + I)^{-1} \hat{w}_{LS} \to \|\hat{w}_{RR}\|_2 \leq \|\hat{w}_{LS}\|_2$

$\hat{w}_{RR} = (\lambda VS^{-2}V^T + VIV^T)^{-1} \hat{w}_{LS} = (V(\lambda S^{-2} + I)V^T)^{-1} \hat{w}_{LS}$

$= (V^T(\lambda S^{-2} + I)V)^{-1} = (V^T)^{-1}(\lambda S^{-2} + I)^{-1} V^{-1} \hat{w}_{LS}$

$\hat{w}_{RR} = V(\lambda S^{-2} + I)^{-1} V^T \hat{w}_{LS} = VMV^T \hat{w}_{LS} \quad M = (\lambda S^{-2} + I)^{-1}$

$M_{ii} = \frac{1}{\lambda S_{ii}^{-2} + 1} = \frac{S_{ii}^2}{\lambda + S_{ii}^2}$

$\hat{w}_{RR} = VMV^T VS^{-1}U^T y = V(\lambda S^{-2} + I)^{-1} S^{-1} U^T y$

$= V(\lambda S^{-1} + S)^{-1} U^T y$

(OII): $\hat{w}_{RR} = VS_\lambda^{-1} U^T y \quad S_\lambda^{-1} = (\lambda S^{-1} + S)^{-1} \quad S_{\lambda ii}^{-1} = \left(\frac{\lambda}{S_{ii}} + S_{ii}\right)^{-1} = \left(\frac{\lambda + S_{ii}^2}{S_{ii}}\right)^{-1}$

(OII): $\hat{w}_{LS} = VS^{-1}U^T y \quad \Rightarrow S_{\lambda ii}^{-1} = \left(\frac{S_{ii}}{\lambda + S_{ii}^2}\right)$

(i)

- $\lambda = 0$: $S_\lambda^{-1} = S^{-1} \Rightarrow \hat{w}_{RR} = \hat{w}_{LS}$

- $\lambda=0$, $s_{dd}\to 0 \Rightarrow \frac{s_{dd}}{\lambda+s_{dd}^2}\to\infty$ (ii)
- $\lambda>0$, $s_{dd}\to 0 \Rightarrow \frac{s_{dd}}{\lambda+s_{dd}^2}\to 0$ (iii)

} Regularisation hyperparameter non-negative ⇒ highly correlated $X_{i,j}$ do not cause $\mathrm{Var}[\hat{w}]$ to blow up.

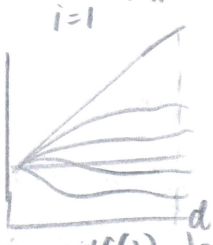
- Ridge regression as least squares: $\underline{y} \approx \hat{X}\underline{w}$

$$\begin{matrix}\\ d\end{matrix}\left\{\begin{bmatrix} y \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0\end{bmatrix}\right. \approx \underbrace{\begin{bmatrix} & X & \\ \sqrt{\lambda} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda}\end{bmatrix}}_{d}\left.\begin{matrix}\\ d\end{matrix}\right\}\begin{bmatrix} w_1 \\ \vdots \\ w_d\end{bmatrix}$$

- Evaluating $\hat{w}_{LS}$ for this problem yields $\hat{w}_{RR}$ of original problem
- Assume std i.e. $\underline{w}\in\mathbb{R}^d$

$$(\underline{y}-\hat{X}\underline{w})^T(\underline{y}-\hat{X}\underline{w}) = (\underline{y}-\underline{X}\underline{w})^T(\underline{y}-\underline{X}\underline{w}) + (\sqrt{\lambda}\underline{w})^T(\sqrt{\lambda}\underline{w})$$
$$= \|\underline{y}-\underline{X}\underline{w}\|_2^2 + \lambda\|\underline{w}\|_2^2$$

- Selecting λ (df plot): $df(\lambda) = \mathrm{trace}\left[\underline{X}(\underline{X}^T\underline{X}+\lambda\underline{I})^{-1}\underline{X}^T\right] = \sum_{i=1}^{d}\frac{s_{ii}^2}{\lambda+s_{ii}^2} = \sum_{i=1}^{d} s_{\lambda ii}^{-1}$
- Plot of $\underline{w}$ against $df(\lambda)$; $\lambda\to\infty$; $df(\lambda)\to 0$ $\lambda\to 0$ $df(\lambda)\to d$



Key equations

- Bias and variance of ridge and ls estimators:
- $\mathbb{E}[\hat{w}_{LS}] = \underline{w}$; $\mathrm{Var}[\hat{w}_{LS}] = \sigma^2(\underline{X}^T\underline{X})^{-1}$ (holds for $\hat{w}_{ML}$ under G-A.)
- $\mathbb{E}[\hat{w}_{RR}] = (\lambda\underline{I}+\underline{X}^T\underline{X})^{-1}\underline{X}^T\underline{X}\underline{w}$; $\mathrm{Var}[\hat{w}_{RR}] = \sigma^2\underline{Z}(\underline{X}^T\underline{X})^{-1}\underline{Z}^T$ $\underline{Z} = (\underline{I}+\lambda(\underline{X}^T\underline{X})^{-1})^{-1}$
- Note convergence of expect. of estimators 1st and 2nd moments under $\lambda=0$!
- $\mathbb{E}[\hat{w}_{RR}/\hat{w}_{LS}]$ and $\mathrm{Var}[\hat{w}_{RR}/\hat{w}_{LS}]$ → insight into how well we can estimate true $\underline{w}$ when our model assumption $y\sim N(\underline{X}\underline{w},\sigma^2\underline{I})$ is correct
- $\hat{w}_{LS}$ - unbiased, high var
- $\hat{w}_{RR}$ - biased, lower variance
- Generalisation: (x_0,y_0) - new data; x_0 measured, y_0 missing

$y_0 = \underline{x}_0^T w_{LS}$ (LS); $y_0 = \underline{x}_0^T w_{RR}$ (RR); $y_0 = \underline{x}_0^T\hat{\underline{w}}$ (general)

- Bias-variance (LR): Expected sq. error of prediction: (generalisation error)

(G.A.) (I) $\mathbb{E}[(y_0-\underline{x}_0^T\hat{\underline{w}})^2|\underline{X},\underline{x}_0] = \int_{\mathbb{R}}\int_{\mathbb{R}^n}(y_0-\underline{x}_0^T\hat{\underline{w}})^2 p(y|\underline{X},\underline{w})\,p(y_0|\underline{x}_0,\underline{w})\,dy\,dy_0$

(I): $\mathbb{E}[(y_0-\underline{x}_0^T\hat{\underline{w}})^2|\underline{X},\underline{x}_0] = \mathbb{E}_{p(y_0|x_0,\underline{w})}\left[\mathbb{E}_{p(y|\underline{X},\underline{w})}\left[(y_0-\underline{x}_0^T\hat{\underline{w}})^2|x_0\right]\,\middle|\,\underline{X}\right]$

(II): $\mathbb{E}[(y_0-\underline{x}_0^T\hat{\underline{w}})^2|\underline{X},\underline{x}_0] = \sigma^2 + \underline{x}_0^T[\underline{w}-\mathbb{E}[\hat{\underline{w}}]][\underline{w}-\mathbb{E}[\hat{\underline{w}}]]^T\underline{x}_0 + \underline{x}_0^T\mathrm{Var}[\hat{\underline{w}}]\underline{x}_0$

(i) noise (can't ctrl) (ii) sq. bias (parameter est. prox to parameter on avr) (iii) variance (sensitivity of pred to data)

(II) derived via:-

(A) independence of y_0 and $\hat{w}$ → $\mathbb{E}[y_0\hat{w}] = \mathbb{E}[y_0]\mathbb{E}[\hat{w}]$

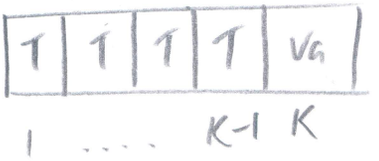
(B) $\mathbb{E}[\hat{w}\hat{w}^T] = \mathrm{Var}[\hat{w}] + \mathbb{E}[\hat{w}]\mathbb{E}[\hat{w}]^T$ (C) $\mathbb{E}[y_0^2] = \sigma^2 + (\underline{x}_0^T\underline{w})^2$ - analogous to (II)

Bias variance (General)
- For model $y = f(\underline{x}; \underline{w}) + \epsilon$ $\mathbb{E}[\epsilon] = 0$ $\text{Var}[\epsilon] = \sigma^2$
- Approximate f by minimising generic loss $\hat{f} = \underset{f}{\text{argmin}}\, L_f$
- Apply $\hat{f}$ to new data $\rightarrow y_0 \approx \hat{f}(x_0) \equiv \hat{f}_0$

Integrate/take exp over all $(y, \underline{X}, \underline{x}_0, y_0)$:-

(BV): $\mathbb{E}[(y_0 - \hat{f}_0)^2] = \sigma^2 + (f_0 - \mathbb{E}[\hat{f}_0])^2 + \text{Var}[\hat{f}_0]$ (note similarity of structure)
noise, sq. bias, variance

X-valid:



K-fold X-valid:
- split data → K groups
- est. model on (K-1) groups, predict K^{th} group (held out)
- do K times
- evaluate perf. using cumul. pred.

(A)

- setting $\lambda \rightarrow$ run (A) for $\lambda_0, \lambda_1, \lambda_2, \dots$,
- select best performing λ
- **Bayes rule** → all 3 variants + probability rules should be instinctive
- **Bridging**: $\hat{w}_{LS} \Leftrightarrow \hat{w}_{ML}$ under G.A. $\underline{y} \sim N(\underline{X}\underline{w}, \sigma^2 \underline{I})$ [or $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$]
- **Ridge regd Bayesian**: likelihood: $\underline{y} \sim N(\underline{X}\underline{w}, \sigma^2 \underline{I})$ (cond. likelihood)
  prior on $\underline{w}$: $\underline{w} \sim N(0, \lambda^{-1}\underline{I}) \rightarrow p(\underline{w}) = \left(\frac{\lambda}{2\pi}\right)^{d/2} e^{-\frac{\lambda}{2}\underline{w}^T\underline{w}}$
- **MAP estimation**:- most probable value under posterior $p(\underline{w}|\underline{y}, \underline{X})$

$$\hat{w}_{MAP} = \underset{\underline{w}}{\text{argmax}}\, \ln p(\underline{w}|\underline{y}, \underline{X})$$

$$= \underset{\underline{w}}{\text{argmax}}\, \ln \left(\frac{p(\underline{y}|\underline{w}, \underline{X})\, p(\underline{w})}{p(\underline{y}|\underline{X})} \right) \quad (p(\underline{y}|\underline{X}) \text{ now constant}) = \underset{\underline{w}}{\text{argmax}} \left(\ln p(\underline{y}|\underline{w}, \underline{X}) + \ln p(\underline{w}) \right)$$

under G.A and G. prior: $\hat{w}_{MAP} = \underset{\underline{w}}{\text{argmax}} -\frac{1}{2\sigma^2}\|\underline{y} - \underline{X}\underline{w}\|_2^2 - \lambda \|\underline{w}\|_2^2 = \underset{\underline{w}}{\text{argmin}} \|\underline{y} - \underline{X}\underline{w}\|_2^2 + \lambda \|\underline{w}\|_2^2$

- **MAP estimate** $\nabla_{\underline{w}} L \Rightarrow \hat{w}_{MAP} = (\lambda\sigma^2 \underline{I} + \underline{X}^T\underline{X})^{-1}\underline{X}^T\underline{y}$
- **MAP estimator and ridge estimator**: $\hat{w}_{MAP} = \hat{w}_{RR}$ with $\lambda_{new} = \lambda\sigma^2$
- **Key relations**: $\hat{w}_{LS} \Leftrightarrow \hat{w}_{ML}$ under Gaussian likelihood $\underline{y} \sim N(\underline{X}\underline{w}, \sigma^2\underline{I})$
  $\hat{w}_{RR} \Leftrightarrow \hat{w}_{MAP}$ under ——— " ——— and Gaussian prior $\underline{w} \sim N(0, \lambda^{-1}\underline{I})$
- RR maximises posterior; LS maximises likelihood

LS Key equations
- **Bayesian LR**: Posterior, predictive, model comparison
- **MAP and ML**: maximising posterior of $\underline{w}$ $\Leftrightarrow$ maximisation of joint likelihood (under G.A. and Gaussian priors)
- ML - only conditional likelihood $p(\underline{y}|\underline{w}, \underline{X})$ considered
  MAP - Accounts for model prior through joint likelihood $p(\underline{y}, \underline{w}|\underline{X}) = p(\underline{y}|\underline{w}, \underline{X})p(\underline{w})$

uncertainty; not just point est ($\hat{w}_{MAP}$ and $\hat{w}_{ML}$) accounts for its

- **Posterior distr.:** $p(w|y,X) = \frac{p(y|w,X)p(w)}{\int_{\mathbb{R}^d} p(y|w,X)p(w)\,dw}$ $\propto p(y|w,X)p(w)$ (i)

- **Posterior updating:** prior $p(w)$ $w \sim N(0, \lambda^{-1}I)$ → likelihood $p(y|w,X)$ $y \sim N(Xw, \sigma^2 I)$ → posterior $p(w|y,X)$

- **Conjugate priors:** Gaussian prior, likelihood → Gaussian posterior (analytic)
- via (i), completion of square/ equating coeff in generalised $p(w|\mu,\Sigma)$ → $p(w|y,X) = \frac{1}{z} e^{-\frac{1}{2}(w^T(\lambda I + \sigma^{-2}X^TX)^{-1}w - 2w^TX^Ty\sigma^{-2})}$

- **Posterior distrib.:** (iii): $p(w|y,X) = N(w|\mu,\Sigma)$ $\mu = (\lambda\sigma^2 I + X^TX)^{-1}X^Ty$ ($\hat{w}_{MAP}$ with $\lambda' = \lambda\sigma^2$)
- $\mu = \hat{w}_{MAP}$ with redefined λ
- Σ - uncertainty about w (like $Var[\hat{w}_{LS}]$ and $Var[\hat{w}_{RR}]$) $\Sigma = (\lambda I + \sigma^{-2}X^TX)^{-1}$
- but this is full distribution (densities, probabilities)

Posterior uses: ① marginal $w_i \sim N(\mu_i, \Sigma_{ii})$ → $w_i > 0$ or $w_i < 0$? ③ Predictive (point predictions)
② Joint marginal $\begin{bmatrix} w_i \\ w_j \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_i \\ \mu_j \end{bmatrix}, \begin{bmatrix} \Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ji} & \Sigma_{jj} \end{bmatrix}\right)$

Predictive distribution: - New pair x_0 measured, y_0 unknown, $\hat{y}_0 = x_0^T\hat{w}_{LS}$ or $x_0^T\hat{w}_{RR}$

(iii): $p(y_0|x_0,y,X) = \int_{\mathbb{R}^d} p(y_0,w|x_0,y,X)\,dw \overset{(ii)}{=} \int_{\mathbb{R}^d} p(y_0|x_0,w)p(w|y,X)\,dw$ · $\mathbb{E}[y_0] = x_0^T\hat{w}_{MAP}$

(ii): $p(y_0|x_0,y,X) = N(y_0|\mu_0,\sigma_0^2)$ $\mu_0 = x_0^T\mu$ and $\sigma_0^2 = \sigma^2 + x_0^T\Sigma x_0$

(under G.A, Gaussian prior) (ii) - y_0 c.i of y,X → $p(y_0|w,x_0,y,X) = p(y_0|w,x_0)$ (as $y_0 \sim N(x_0^Tw, \sigma^2)$)
w c.i of x_0 → $p(w|x_0,y,X) = p(w|y,X)$

Batch posterior: $p(w|y_0,x_0,y,X) \propto p(y_0|w,x_0)\,p(w|y,X)$
prior → posterior (full posterior over all data) (likelihood new y_0 given w) (new prior on w post (x_0,y_0) obs) / (posterior on old data (y,X))

Posterior sequential updating: (I) $p(w|y,X) = N(w|\mu,\Sigma)$ $\mu = (\lambda\sigma^2 I + X^TX)^{-1}X^Ty$
$\Sigma = (\lambda I + \sigma^{-2}X^TX)^{-1}$

(II) observe (x_0,y_0)

(III) $p(w|y_0,x_0,y,X) = N(w|\mu,\Sigma)$

(iii) $\mu = \left(\lambda\sigma^2 I + (x_0x_0^T + \sum_{i=1}^{n} x_ix_i^T)\right)^{-1}$
$\Sigma = \left(\lambda I + \sigma^{-2}\left(x_0x_0^T + \sum_{i=1}^{n} x_ix_i^T\right)\right)^{-1}\left(x_0y_0 + \sum_{i=1}^{n} x_iy_i\right)$

Updating $\hat{w}$ + predictions for new y_0: (I) Form predictive $p(y_0|x_0,y,X)$ → obs. y_0
(II) update posterior $p(w|y_0,x_0,y,X)$ with rank 1 update

Active learning (Heuristic): 1) Predictive $p(y_0|x_0, y, X)$ for all $x_0 \in D$ that have no y_0 response
2) Select x_0 for which $\sigma_0^2 = \sigma^2 + x_0^T \Sigma x_0$ is greatest, measure y_0
3) Update post. of $p(w|y, X)$ via $y_{new} \leftarrow (y, y_0)$ $X_{new} \leftarrow (X, x_0)$
4) Return to #1

Differential entropy: Heuristic optimises diff. entropy $H(p) = -\int p(z) \ln p(z) dz$
for continuous p.d. $p(z)$ [H is a functional]

- $H(p) > 0$, $H(p) \to \infty$ (more uncertainty) $H(p) < 0$ $H(p) \to -\infty$ (less uncertainty)

MVG diff. entropy: $p(z) = N(w|\mu, \Sigma) \Rightarrow H(p) = H(N(w|\mu, \Sigma)) = \frac{d}{2} \ln(2\pi e |\Sigma|)$

- H depends on i) dimensionality of distri. of w (d)
ii) determinant of posterior covariance of matrix of w ($|\Sigma|$)

- How does post. cov $|\Sigma|$ change with new measurements, and effect on $H(p)$?

- (ITT): $H_{post} = H_{prior} - \frac{d}{2} \ln(\underbrace{1 + \sigma^{-2}(x_0^T \Sigma x_0)}_{(i)})$ (i) Update is a function of x_0 and Σ, because post cov does not depend on measured y_0

- The x_0 selected that $\max(\sigma_0^2)$ also minimises H_{post}. post. (myopically/greedily)
(pred. variance) (uncertainty)

- **Model selection**: selection of λ (hyperparam.) (prior precision)
(Bayesian evidence max.) i) X-validation
ii) evidence max. $\longrightarrow \hat{\lambda} = \arg\max_{\lambda} \ln p(y|X, \lambda)$ log-likelihood of data having integrated out param w

- $p(w|y, X, \lambda) = \frac{p(y|w, X) p(w|\lambda)}{p(y|X, \lambda)}$ $p(y|X, \lambda) = N(y|0, \sigma^2 I + \lambda^{-1} X X^T)$ (*)

- Type I ML: maximise (cond.) likelihood over man param.
- Type II ML: integrate out man param w, maximise over hyperparam λ (E.B.)
- Require solution to $\int_{\mathbb{R}^d} p(y|w, X) p(w|\lambda) dw$ (analytic), or iteratively (Bishop)
(i) re-estimation
(ii) E-M

Key equations (*)

- underdetermined i.e. $(d >> n) \to (X^T X)^{-1}$ not full-rank
- Infinite solutions w satisfying $y = Xw$

Least norm solution: $\hat{w}_{in} = X^T (X X^T)^{-1} y \Rightarrow X \hat{w}_{in} = X X^T (X X^T)^{-1} y$

Generation of sol.: Append non-zero $\partial \neq 0 \in \mathbb{R}^d$ in null-space of X
$\partial \in N(X) \Rightarrow X\partial = 0$ $\partial \neq 0 \Rightarrow X(\hat{w}_{in} + \partial) = X\hat{w}_{in} + X\partial = y + 0$

(OTT): $\hat{w}_{in}$ is solution to overdetermined problem with least l_2 norm
i) via analysis - $\hat{w}_{in} = \arg\min_w \|w\|_2^2$ st. $Xw = y$
ii) via Lagrange multipliers

Least norm proof(3): Define arbitrary sol. $\hat{w}_{sol}$ to $y = X\hat{w} \Rightarrow X\hat{w}_{sol} - X\hat{w}_{LN} = 0$

- $(\hat{w}_{sol} - \hat{w}_{LN})$ orthogonal to $\hat{w}_{LN}$
- L_2 norm decomposition: $\|\hat{w}_{sol}\|_2^2 \to \|\hat{w}_{sol} - \hat{w}_{LN}\|_2^2 + \|\hat{w}_{LN}\|_2^2 > \|w_{LN}\|_2^2$
- Equality when $\hat{w}_{sol} = \hat{w}_{LN}$

Sparse L_1: · LS, RR not suitable for moderate / high-dim problems } treat all dim. equally without favouring subsets
· Feature sel may $\Rightarrow$ few features important for prediction

Ridge reg analysis: $L = \sum_{i=1}^{n} (y_i - f(x_i; w))^2 + \lambda \|w\|_2^2$ $f(x_i, w) = x_i^T w$

- decompose as loss = goodness of fit (in-sample) + penalty term (expresses preferences about w)

L_2-norm Quadratic penalty: $\lambda \|w\|_2^2 = \lambda \sum_{i=1}^{d} w_i^2$; individual w_j (dimension) $\Rightarrow \frac{\partial \lambda \sum_{i=1}^{d} w_i^2}{\partial w_j} = 2 w_j \lambda$

- $\frac{\partial \lambda \|w\|_2^2}{\partial w_j}$ is function of w_j and magnitude, start point

(50): favours vectors $\hat{w}$ with small entries of similar size
· L_2 norm penalty $\not\Rightarrow$ sparsity encourages similar size estimates $\hat{w}_j$ elements
· select a subset of d features, switch off rest. ↘

Sparsity/feature subsets: each w_j corresponds to dim of x:
$w_j = 0$ $f(x; w) = x^T w = w_1 x_1 + \ldots + 0 \cdot x_j + \ldots + w_d x_d$

· Feature sel: Find w that predicts well and has small no. of non-zero entries

Sparsity: w where most w_js are 0 (sparse)

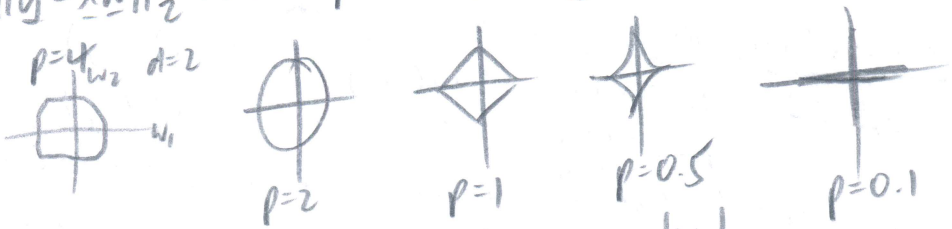
· Linear penalty (sparsity): Achieved with linear penalty term: w_R large, other w_js small and non-zero (close to zero) } check

· Quadratic penalty: Favour entries with similar size, push w_R to average value

Lasso/L_1 regularised reg: $\hat{w}_{lasso} = \underset{w}{\operatorname{argmin}} \|y - Xw\|_2^2 + \lambda \|w\|_1^2$ $\|w\|_1^2 = \sum_{j=1}^{d} |w_j|$

L_p regression: $\hat{w}_{Lp} = \underset{w}{\operatorname{argmin}} \|y - Xw\|_2^2 + \lambda \|w\|_p^p$ $\|w\|_p = \sum_{j=1}^{d} (|w_j|^p)^{1/p}$ $0 \le p \le \infty$

· L_p penalisation (level sets)



· $\|\cdot\|_p$ behaviour: $p=\infty$, $\|\cdot\|_\infty$ → largest abs. entry $\|w\|_\infty = \max_j |w_j|$
$p>2$, $\|\cdot\|_{p>2}$ → focus on large entries
$p=2$, $\|\cdot\|_2$ → large entries expensive
$p=1$, $\|\cdot\|_1$ → sparsity; non-convex contour-sets ↕
$p<1$, $\|\cdot\|_{p<1}$ → sparsity encouraged (LASSO)
$p \to 0$ $\|\cdot\|_0$ → $\sum_j \mathbb{I}(w_j = 0)$ records if entry non-zero.

L6-Lp regul. reg solutions: LS → Analytic if (X^TX) invertible
RR → Analytic sol always
LASSO → Numerical opt. via convex opt. (global)
ℓ_p ($p \geq 1, p \neq 2$) → convex optim. → global
($p < 1$) → approximate local solns only via iterative methods

L7 Key equations

- **Classification:** Input: Measurements $(\underline{x}_1, \dots, \underline{x}_n)$ in input space $\underline{x}_i \in \mathbb{R}^d$
  Output: Discete output space $Y = \{+1, -1\}$ or $\{0, 1\}$ or $Y = \{1, \dots, K\}$
- Define a function f (classifier) to map input $\underline{x}$ to class y: $y = f(\underline{x})$
- Est. Define classifier and estimate its parameters
- **NN-classifier:** Given $(\underline{x}_1, y_1), (\underline{x}_2, y_2), \dots, (\underline{x}_n, y_n)$ construct classifie $\hat{f}(\underline{x}) \to y$ (A)
- For a new input $\underline{x}$ not in training data:-
(I) Let $\underline{x}_i$ be point in $(\underline{x}_1, \dots, \underline{x}_n)$ closest to $\underline{x}$
(II) Return label y_i
- **Distance metric:** For data in $\mathbb{R}^d$ → ℓ_2 norm/Euclidean: $\|\underline{u} - \underline{v}\|_2 = \left(\sum_{i=1}^d (u_i - v_i)^2\right)^{1/2}$
  ↳ ℓ_p $p \in [1, \infty]$: $\|\underline{u} - \underline{v}\|_p = \left(\sum_{i=1}^d (u_i - v_i)^p\right)^{1/p}$
- Non $\mathbb{R}^d$ data: edit/correl distance
- **Training error:** $err(\hat{f}, S)$
- **Test error:** $err(\hat{f}, T)$
- **K-NN classifier:** (A)
- For new input $\underline{x}$:-
(I) Return K points closest to $\underline{x}$, indexed $x_{i1}, x_{i2}, \dots, x_{ik}$
(II) Return majority vote of $y_{i1}, y_{i2}, \dots, y_{ik}$
  - Break ties in steps arbitrarily
  - OCR example → sweep for various K, select best
- **Effects of K:** smaller K - smaller training error
  large K - predictions more stable (majority vote)
- **Decision boundary:** Region of problem space in which output label of classifie changes
- **Classifie quality:** {Prediction accuracy :- $P(f(\underline{x}) = y)$
  Prediction error :- $P(f(x) \neq y)$
- Requires a distribution P ove space of labelled examples generating data
- $(\underline{x}_i, y_i) \overset{iid}{\sim} P \quad \forall n = 1, \dots, N$ (drawn iid from unknown ground truth distri., joint distribution ove covariates and label)
- **Statistical learning theory:** IID is a uniformity of nature assumption
  Theoretical guarantees in SLT by Vapnik et al.
- **Optimal classifiers:** For any classifier $f: X \to Y$, prediction error

$$P(f(X) \neq Y) = \mathbb{E}\left[\mathbb{I}(f(X) \neq Y)\right] = \mathbb{E}_{P(X)}\left[\mathbb{E}_{P(Y|X)}\left[\mathbb{I}(f(X) \neq Y) \mid X\right]\right]$$

(*)

- minimise $P(f(X) \neq Y)$ via $\min (*) = \sum_{y \in \mathcal{Y}} P(Y=y|X) \cdot \mathbb{I}(f(X) \neq Y)$ (†)
- (*) minimised for particular $\underline{x} \in \mathcal{X}$ when:-
- $f(\underline{x}) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, P(Y=y|X=\underline{x})$ - Bayes classifier, f^* property → lowest prediction error amongst all classifiers
- (*): Assign label to $\underline{x}$ to be most probable label according to distribution specified by $P \Rightarrow \mathbb{I}(f(X) \neq Y) = 0$ for maximum value $P(Y=y|X=\underline{x})$

Optimal classifier reserv: labelled iid data from P; for every $x \in \mathcal{X}$, predict label to be most probable conditioning on $\underline{x}$, according to that distri.

- We do not know P, use $P(Y=y, X=x)$ to approx.

Bayes classifier: under $(X,Y) \overset{iid}{\sim} P$, optimal classifier → $f^*(\underline{x}) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, P(Y=y|\underline{X}=\underline{x})$

Equivalent to $f^*(\underline{x}) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, P(Y=y) P(X=\underline{x}|Y=y)$

Why? $P(f(X) \neq Y) = \sum_{y \in \mathcal{Y}} P(Y=y|X=\underline{x}) \cdot \mathbb{I}(f(\underline{x} \neq y)$; set $\mathbb{I}(f(X) \neq Y) = 0$ for max $P(Y=y|X=\underline{x})$

- Then sum up $(K-1)$ smallest probabilities according to $P(Y=y|X=\underline{x})$ (easier if think of K classes)

→ class prior $P(Y=y)/P(Y=R)$; class condition distri $P(X=\underline{x}|Y=k)$ - unknown; approximate

Gaussian CCDs: $y \in \{0,1\}$ $X \in \mathbb{R}^1$ $P(X,Y)$ given by (generatively):-

$P(Y=R) = \pi_R$ $y \in \{k=0, k=1\}$; $P(X|Y=k) = p_R(\underline{x}) = N(\underline{x}|\mu_R, \sigma_R^2)$ → $p_0(\underline{x})$, → $p_1(\underline{x})$

Example #2 Bayes classif: $f^*(\underline{x}) = \underset{y \in \{0,1\}}{\operatorname{argmax}}\, P(X=x|Y=R) P(Y=k)$

(Gaussian CCD)

$f^*(x) \begin{cases} 1 \text{ if } x > \frac{1}{2} \\ 0 \text{ otherwise} \end{cases}$ (Bayes class.) yields ← $\begin{cases} 1 \text{ if } \frac{\pi_1}{\sigma_1} \exp\left[-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right] > \frac{\pi_0}{\sigma_0} \exp\left[-\frac{(x-\mu_0)}{2\sigma_0^2}\right] \\ 0 \text{ otherwise} \end{cases}$

Example #3 MVGs: $y \in \{0,1\}$ $x \in \mathbb{R}^2$ $P(X,Y)$ given by:-

$P(Y=k) = \pi_R$; $P(X|Y=k) = p_R(\underline{x}) = N(\underline{x}|\mu_R, \Sigma_R)$ distinct Σ_0 and Σ_1 (Gaussian CCDs MVG)

- shared $\Sigma_0 = \Sigma_1$: Bayes classifier → linear decision boundary
- $\Sigma_0 \neq \Sigma_1$: —"— → quadratic —"—
- Bayes classifier → complicat. if complicat. CCDs

Plug-in classifiers: (BUT) Bayes classifier → smallest predi error $P(f(X) \neq Y)$

- cannot construct without knowing $P(X,Y)$ → $P(Y=R) = ?$ $P(Y=k|X=\underline{x}) = ?$ $P(X=\underline{x}|Y=k) = ?$
- Only labelled iid examples from P
- Ingneral, CCDs unknown, approx.

Plug-in classifiers: use available data to approx $P(Y=R)$ and $P(X=\underline{x}|Y=R)$

Example #4: $X \in \mathbb{R}^d$ $y \in \{1, \dots, K\}$, Estimate Bayes classifier via MLE

(MVG CCDs, plug-in)

class priors: $\hat{\pi}_R = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(y_i = R)$

class cond distrs $P(X=\underline{x}, Y=k) \rightarrow N(\underline{x}|\mu_R, \Sigma_R)$ empirical mean of class y

MLE CCD param est: $\hat{\mu}_R = \frac{1}{n_R} \sum_{i=1}^N \mathbb{I}(y_i=k) x_i$ $\hat{\Sigma}_R = \frac{1}{n_R} \sum_{i=1}^N \mathbb{I}(y_i=k)(x_i - \hat{\mu}_R)(x_i - \hat{\mu}_R)^T$ cov.

Example #4 plug-in: $\hat{f}(\underline{x}) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \left[\hat{\pi}_k |\hat{\Sigma}_k|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\underline{x} - \hat{\mu}_k) \hat{\Sigma}_k^{-1} (\underline{x} - \hat{\mu}_k)^T \right\} \right]$

- Eval. to RHS for new $\underline{x}$ for each of $y = 1, \ldots, K$ classes; select $y = k$: argmax ...

Spam filtering (bag of words): (Example #5) Let $\underline{x}$ be a vector of word counts with indexed element corresponding to word ($j \to$ cat) $x(j) = 3$

Naive Bayes assumption: Treat dimensions of X as conditionally independent given $Y = k$ (label) i.e. $P(X = \underline{x} | Y = y) = \prod_{j=1}^{d} p_j(x(j) | Y = k)$

- spam → word correlations ignored
- distribution of 'histogram of word counts' given class assignment is independent

Spam filtering: estimation: class prior $P(Y = k)$ estimated from training data

$$\hat{P}(Y = k) = \frac{\#\text{ obs in class } k}{\#\text{ obs}}$$

Class-cond. distribution: CCD: $P(X = \underline{x} | Y = k) = \prod_j p_j(x(j) | Y = k) = \prod_j P_0\left(x(j) | \lambda_j^{(k)}\right)$

- Estimation of Poisson param via MLE:

$$\hat{\lambda}_j^{(k)} = \frac{\#\text{ unique uses of word } j \text{ in obs from class } k}{\#\text{ obs in class } k} \quad (*)$$

- walk through intuition

- $P_0(x(j) | \lambda_j^{(k)})$: Given an email from class k, no. of word occurrences is Poisson distri with $\lambda_j^{(k)}$ for class k; every j^{th} word has poisson param associated with it, dependent on class.
- $\hat{\lambda}_j^{(k)}$ (*) has to be estimated for all words indexed by j in vocabulary; and all classes k.

18 - Key equations:

Binary classification: $\underline{x}_i \in \mathbb{R}^d$ $y_i \in \{\pm 1\}$ $\underline{\theta}$ - parameter vector
classifier f: $y_i = f(\underline{x}_i; \underline{\theta})$; $f: \mathbb{R}^d \to \{-1, +1\}$

Bayes classific. frame: $\underline{\theta} \to$ i) class prior distr on y
→ ii) class dependent distri on $\underline{x}$ parameters

Linear classific. frame: Prediction linear in $\underline{\theta}$, intersection of Bayes class cd linear class

Bayes classifier (log-odds): Predict class of $\underline{x}$ to be most probable label given model and training data $(\underline{x}_1, y_1) \ldots (\underline{x}_n, y_n)$

(Binary case)

- declare class $y = 1$ if $p(\underline{x}|y=1)\underbrace{p(y=1)}_{\pi_1} > p(\underline{x}|y=0)\underbrace{p(y=0)}_{\pi_0}$

log-odds form: IF $\ln\left[\frac{p(\underline{x}|y=1)p(y=1)}{p(\underline{x}|y=0)p(y=0)}\right] > 0$ declare class $\underline{x} \to y = 1$

ELSE ——— = ——— → $y = 0$

LDA:- (Gaussian shared covariance Bayes classifie)
- Log-odds for:- $p(y=0)=\pi_0$, $p(y=1)=\pi_1$, $p(\underline{x}|y=R)=N(\underline{x}|\underline{\mu}_R, \underline{\Sigma})$
- Note shared $\underline{\Sigma}$
- Log odds: $\ln\left[\frac{p(\underline{x}|y=1)p(y=1)}{p(\underline{x}|y=0)p(y=0)}\right] = \ln\left(\frac{\pi_1}{\pi_0}\right) - \frac{1}{2}(\underline{\mu}_1+\underline{\mu}_0)^T\underline{\Sigma}^{-1}(\underline{\mu}_1-\underline{\mu}_0) + \underline{x}^T\underline{\Sigma}^{-1}(\underline{\mu}_1-\underline{\mu}_0)$
  $= w_0 \qquad\qquad + \underline{x}^T\underline{w}$ ($\underline{w} = \underline{\Sigma}^{-1}(\underline{\mu}_1-\underline{\mu}_0)$)

- LDA-Bayes classifier (relation): decision rule for Bayes classifier: $f(\underline{x}) = \text{sign}(\underline{x}^T\underline{w} + w_0)$
- Note Bayes class. → linear decision boundary in $\mathbb{R}^d$ under shared covariance of CCDs for both classes
- Bayes classifie - one instance of a linear classifier with $\begin{cases} w_0 = \ln\left(\frac{\pi_1}{\pi_0}\right) - \frac{1}{2}(\underline{\mu}_1+\underline{\mu}_0)^T\underline{\Sigma}^{-1}(\underline{\mu}_1-\underline{\mu}_0) \\ \underline{w} = \underline{\Sigma}^{-1}(\underline{\mu}_1-\underline{\mu}_0) \end{cases}$
- Assuming single Gaussian and shared covariance

Binary linear classifier: A function of form: $f(x) = \text{sign}(\underline{x}^T\underline{w} + w_0)$ $\underline{w} \in \mathbb{R}^d$ $w_0 \in \mathbb{R}$
- Assuming linear separability; estimate $\hat{w}_0$ and $\underline{w}$ from training data of classes

linear separability: Two sets A, B linearly separable if there exists a $(\underline{w}, w_0)$ such that
$\underline{x}^T\underline{w} + w_0 \begin{cases} > 0 \text{ if } x \in A \text{ (class +1)} \\ < 0 \text{ if } x \in B \text{ (class -1)} \end{cases}$ (for a dimensionality $\underline{x} \in \mathbb{R}^d$)

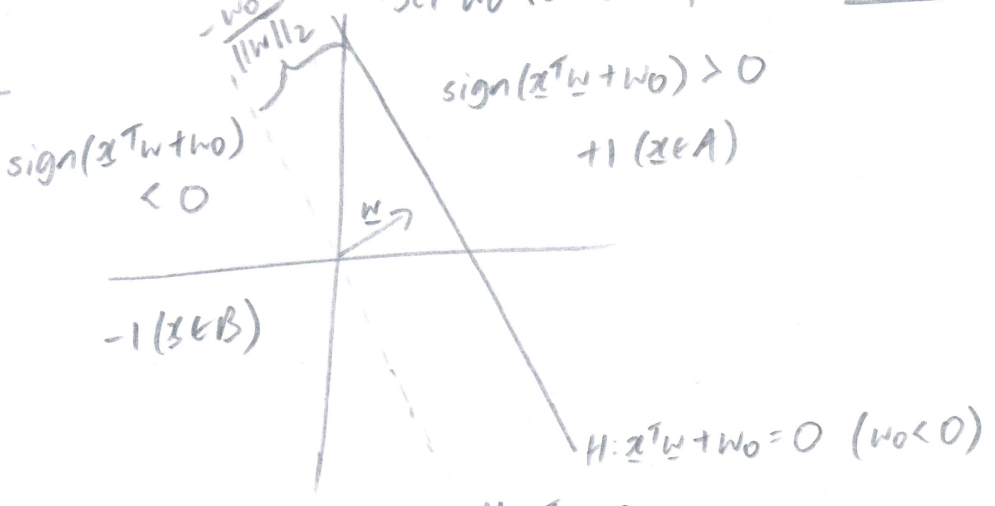
- The pair $(\underline{w}, w_0)$ → affine hyperplane (Hyperplane, affine hyp crucial for geo. intuition)

Hyperplane:
- A hyperplane in $\mathbb{R}^d$ is a $\mathbb{R}^{d-1}$ linear subspace (contain origin)
- H hyperplane H represented with vector $\underline{w}$ (param vector)
- $H = \{\underline{x} \in \mathbb{R}^d : \underline{x}^T\underline{w} = 0\}$ (locus of points in $\mathbb{R}^d$ orthogonal to $\underline{w}$)

Affine hyperplane: An affine hyperplane in $\mathbb{R}^d$ is a hyperplane shifted using a scalar w_0
- $H = \{\underline{x} \in \mathbb{R}^d : \underline{x}^T\underline{w} + w_0 = 0\}$
- Set $w_0 > 0$ → shift H in opposite direction of $\underline{w}$
- Set $w_0 < 0$ → shift H in direction of $\underline{w}$

- Geometric illustration:

(?) (*) - intuitions



- Classification: Find $(\hat{\underline{w}}, \hat{w}_0)$ such that $\text{sign}(\underline{x}^T\underline{w}+w_0) > 0$ on all $\underline{x} \in A$ (+1)
  (angle) (shift) $\text{sign}(\underline{x}^T\underline{w}+w_0) < 0$ on all $x \in B$ (-1)
- For given $\underline{w}$, w_0, input $\underline{x}$; $\underline{x}^T\underline{w} + w_0 > 0 \Rightarrow \underline{x}$ is on far side of affine hyper H in direction $\underline{w}$ points.

Polynomial generalisations:

- use linear classifier on $\underline{x} = (x_1, x_2)$ $\underline{x} = (x_1, x_2, x_1^2, x^2)^T$
- Binary L.C. in $\mathbb{R}^4$ → find $\mathbb{R}^3$ hyperplane in $\mathbb{R}^4$ to separate vectors
- D.B. in $\mathbb{R}^4$ projected → $\mathbb{R}^2$ - quadratic in $\mathbb{R}^2$; but linear in $\mathbb{R}^4$
- linear separability in $\mathbb{R}^4$
- Bayes class. (shared covariance) → linear classification
- Bayes class (different covariances) → polynomial classification

QDA: (Gaussian distinct covariance Bayes classifie)

log-odds for $p(y=0)=\pi_0$ $p(y=1)=\pi_1$ $p(\underline{x}|y=k) = N(\underline{x}|\underline{\mu}_k, \underline{\Sigma}_k)$

Note distinct $\underline{\Sigma}_k$

- log-odds: $\ln\left[\frac{p(y=1)p(\underline{x}|y=1)}{p(y=0)p(\underline{x}|y=0)}\right] = \text{const.} + \underline{x}^T(\underline{\Sigma}_1^{-1}\underline{\mu}_1 - \underline{\Sigma}_0^{-1}\underline{\mu}_0) + \underline{x}^T(\underline{\Sigma}_0^{-1}/2 - \underline{\Sigma}_1^{-1}/2)\underline{x}$

  linear in $\underline{x}$; quadratic in $\underline{x}$

- linear in weights

B.C. ⇔ L.C. ⇔ P.L: QDA: $f(\underline{x}) = \text{sign}(\underline{x}^T A \underline{x} + \underline{x}^T b + c)$ linear in A, b, c

$\text{sign}(\underline{x}^T A \underline{x} + \underline{x}^T b + c) > 0$ class 1

$\text{sign}(x^T A x + x^T b + c) < 0$ class 0 or (-1)

- $\underline{x} \in \mathbb{R}^2$; rewrite $\underline{x} \leftarrow (x_1, x_2, 2x_1x_2, x_1^2, x_2^2)$, do linear classification in $\mathbb{R}^5$
- L.C. estimation in $\mathbb{R}^5$ maps to non-linear fn. in $\mathbb{R}^2$

General classifies ⇔ regres: more general classifies $f(\underline{x}) = \text{sign}(\underline{x}^T\underline{w} + w_0)$

- 'Regression':
  1. Define $\underline{y} = (y_1, \dots, y_n)^T$ with y_i = class labels
  2. Append 1s to each $\underline{x}_i$ and construct $\underline{X} = [\underline{x}_1, \dots, \underline{x}_n]^T$ ← w_0 included in $\underline{w}$
  3. Estimate $\hat{\underline{w}}_{LS} = (\underline{X}^T\underline{X})^{-1}\underline{X}^T\underline{y}$
  4. New $\underline{x}_0$, declare $y_0 = \text{sign}(\underline{x}_0^T\underline{w})$

↓

- can use lin reg. as baseline; use with k-NN as baseline
- sensitivity to outliers (least squares classification) → motivates method that is:-

a) robust b) less sensitive to covariates c) focus on hyperplane

Perceptron (Rosenblatt 1958):

- **Linear separability:** Is added $\underline{x}_i$s

  data in $\mathbb{R}^d$ linearly separable if it's possible to find $(\underline{w}, w_0)$ defining class. $y_i = \text{sign}(\underline{x}_i^T\underline{w})$ $\forall i$ with zero-training error
- Issue - there may be infinite no. of classifies meeting this criterion. Is added $(w_0 \to \underline{w})$
- **Perceptron objective:** using linear class. $y = f(\underline{x}) = \text{sign}(\underline{x}^T\underline{w})$ $y \in \{-1, +1\}$
  - minimises $\mathcal{L} = -\sum_{i=1}^{N}(y_i \cdot \underline{x}_i^T\underline{w})\mathbb{I}(y_i \neq \text{sign}(\underline{x}_i^T\underline{w}))$

    over misclassified points
  - $y \in \{-1, +1\} \Rightarrow y_i \cdot \underline{x}_i^T\underline{w}$ $\begin{cases} > 0 & \text{if } y_i = \text{sign}(\underline{x}_i^T\underline{w}) \\ < 0 & \text{if } y_i \neq \text{sign}(\underline{x}_i^T\underline{w}) \end{cases}$

- **Iterative methods for estimating:** min $\mathcal{L}$ via analytic ×
- **Stochastic gradient descent:-** For sufficiently small η

  $\underline{w}' \leftarrow \underline{w} - \eta \nabla_{\underline{w}} \mathcal{L} \Rightarrow \mathcal{L}(\underline{w}') < \mathcal{L}(\underline{w})$

  - update $\underline{w}'$ by moving in small increment in opposite (negative) direction of $\underline{w}$