Idea: $\mathbf{w}^T\mathbf{x} + w_0 = \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}$, where
$\tilde{\mathbf{w}} := [w_1, ..., w_d, w_0]^T$; $\tilde{\mathbf{x}} = [x_1, ..., x_d, 1]^T$

Def: Residual: $r_i = y_i - f(y_i)$; Loss function l;
l^p-loss: $l(r) = |r|^p$; Emp. risk: $\hat{R}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n l(r_i)$

LSR problem: $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{i=1}^n (y_i - \mathbf{w}^T\mathbf{x}_i)^2$

LSR expl. sol.: $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$; $O(nd^2+d^3)$

Gradient Descent (G.D.) $\nabla\hat{R} \Rightarrow O(nd)\log(\frac{1}{\epsilon})$:
`1. Start at an arbitrary` $\mathbf{w}_0 \in \mathbb{R}$,
`2. For t=0,1,2,... do:` $\mathbf{w}_{t+1} = \mathbf{w} - \eta_t\nabla\hat{R}(\mathbf{w}_t)$.
· R convex ⇒ G.S. converges; $l = l^2, \eta_t = \frac{1}{2} \Rightarrow O(t)$

Adaptive step size: (Add step 3. in G.D. above)
1.Line search: 3. $\eta_t^* = \arg\min_{\eta\in[0,\infty)} \hat{R}(\mathbf{w}_t - \eta g_t)$.
2.Bold driver: 3.If $\hat{R}(\mathbf{w}_t) < \hat{R}(\mathbf{w}_t)$: $\eta_t := c_{acc}\eta_{t-1}$,
else $\eta_t := c_{dec}\eta_{t-1}$.

Non-lin. reg.: $f(x) = \sum_{i=1}^d w_i\phi_i(\mathbf{x})$, $\mathcal{B}_H = (\phi_i)_i$

ERM:· LoLN ⇒ $\hat{R}(\mathbf{w}) \xrightarrow{|D|\to\infty} R(\mathbf{w})$ a.s..
· $l = l^2$, supp(D)$<\infty \Rightarrow ||R - \hat{R}|| \to 0$ (in C^0)
· $\mathbb{E}_D[\hat{R}_D(\hat{\mathbf{w}}_D)] \le \mathbb{E}_D[R(\hat{\mathbf{w}}_D)]$ (Pf: Jensen's (swap))

Idea: Use train/val./test sets, reduce general. error
·Optimize $\hat{\mathbf{w}}_{D_{train}} = \arg\min_{\mathbf{w}} \hat{R}_{train}(\mathbf{w})$, but
evaluate $\hat{R}_{test}(\hat{\mathbf{w}}) = \frac{1}{|D_{test}|}\sum_{(\mathbf{x},y)\in D_{test}} (y - \hat{\mathbf{w}}^T\mathbf{x})^2$.
· $\mathbb{E}_{D_{tr.},D_{test}}[\hat{R}_{D_{test}}(\hat{\mathbf{w}}_{D_{tr.}})] = \mathbb{E}_{D_{tr.}}[R(\hat{\mathbf{w}}_{D_{tr.}})]$(iid)
MC/k-fold cross validation (only when D idd):
`1. For candidate model m and i=1,...,k:`
a) Split (train) data: $D = D_{train}^{(i)} \sqcup D_{val}^{(i)}$
b) Train model: $\hat{\mathbf{w}}_{i,m} = \arg\min_{\mathbf{w}} \hat{R}_{train}^{(i)}(\mathbf{w})$
c) Estimate error: $\hat{R}_m^{(i)} = \hat{R}_{val}^{(i)}(\hat{\mathbf{w}}_i)$
2. Select model: $\hat{m} = \arg\min_m \frac{1}{k}\sum_{i=1}^k \hat{R}_m^{(i)}$

k large: Risk overfitting to D_{val}, underfitting to D_{train} and having too little data for training
k small: Higher $O(\cdot)$ but better performance
k=n: LOOCV; in practice often k=5 or k=10

RR prob.: $\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^n (y_i - \mathbf{w}^T\mathbf{x}_i)^2 + \lambda||\mathbf{w}||_2^2$
RR expl. sol.: $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$ (std $\mathbf{X}$!),
$x_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j} | \hat{\mu}_j = \frac{1}{n}\sum_i^n x_{ij} | \hat{\sigma}_j = \frac{1}{n}\sum_i^n (x_{ij} - \hat{\mu}_j)^2$
RRGD:2.For t: $\mathbf{w}_{t+1} = (1-2\lambda\eta_t)\mathbf{w}_t - \eta_t\nabla\hat{R}(\mathbf{w}_t)$

General regularizatoin: $\min_{\mathbf{w}} \hat{R}(\mathbf{w}) + \lambda C(\mathbf{w})$
· Tradeoff: g.o.f. vs. simplicity ($\lambda \gg 0$ higher $O(\cdot)$)
λ choice: CV w. e.g. $m(\lambda), \lambda \in \{10^{-6}, 10^{-5}, .., 10^6\}$

Bin. lin. classifiers: $f(\mathbf{x}) = f_{\mathbf{w}}(\mathbf{x}) = \text{sgn}(\mathbf{w}^T\mathbf{x})$,
$l = l_{0/1}(\mathbf{w}; \mathbf{x}_i, y_i) := 1[y_i \ne f_{\mathbf{w}}(\mathbf{x}_i)]$ (a.e. $\nabla_{\mathbf{w}} = 0$!)
Surrogate losses: $l_P(\mathbf{w}; \mathbf{x}, y) = max(0, -y\mathbf{w}^T\mathbf{x})$,
$l_H(\mathbf{w}; \mathbf{x}, y) = max(0, 1 - y\mathbf{w}^T\mathbf{x})$
GD: 2.For t: $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \sum_{i\in\mathcal{I}_{\mathbf{w}_t}} y_i\mathbf{x}_i$, where
$\mathcal{I}_{\mathbf{w}} = \{i : (\mathbf{x}_i, y_i)$ incorrectly classified by $\mathbf{w}\}$ (inef.!)
Idea: Evaluate only a k pts in $\mathcal{I}_{\mathbf{w}}$ ($k=1 \Rightarrow$ SGD)
SGD:1.Start with arbitrary $\mathbf{w}_0 \in \mathbb{R}^d$
2. For t=0,1,2,.. do:
a) Pick $(\mathbf{x}', y') \in D_{train}$ U-randomly
b) Set $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t\nabla l(\mathbf{w}_t; \mathbf{x}', y')$
Conv: Guar. if $\sum_t \eta_t = \infty$ and $\sum_t \eta_t^2 < \infty$
Minibatch SGD: a) $k>1$ and in b) take ∇l avg
"Mini-batches exploit parallelism, reduce variance"

Perceptron alg (PA): SGD with $l = l_P$ and $\eta_t = 1$
Thm: If data lin. seperable, PA finds lin. separator
SVM: $\min_{\mathbf{w}} \frac{1}{n}\sum l_H(\mathbf{w}; \mathbf{x}_i, y_i) + \lambda||\mathbf{w}||_2^2$; ip:$\eta_t = \frac{1}{\lambda t}$
SGD:b) $\mathbf{w}_{t+1} = (1 - \frac{2\lambda}{n}\eta_t)\mathbf{w} + 1[y_i\mathbf{w}^T\mathbf{x}_i < 1]\eta_t y_i\mathbf{x}_i$

Greedy forward selection: Feat.s $V = \{1, .., d\}$,
feat. selection $S \subseteq V$, CV-Loss $\hat{L}(S)$:
1. Start with $S=\emptyset$ and $E_0 = \infty$
2. For i=1,..,d, do:
a)Find best feature:$s_i = \arg\min_{j\in V\setminus S} \hat{L}(S\cup\{j\})$
b)Compute error: $E_i = \hat{L}(S\cup\{s_i\})$
c)If $E_i > E_{i-1}$ break, else set $S = S\cup\{s_i\}$

Greedy backward selection:(-slower,+dep. feats)
1. Start with $S = V$ and $E_{d+1} = \infty$
2. For i=d,..,1, do:
a)Find best feature:$s_i = \arg\min_{j\in S}\hat{L}(S\setminus\{j\})$
b)Compute error: $E_i = \hat{L}(S\setminus\{s_i\})$
c)If $E_i > E_{i+1}$ break, else set $S = S\setminus\{s_i\}$

Alt: $\hat{\mathbf{w}} = \arg\min \sum l(\mathbf{w}; \mathbf{x}_i, y_i) + \lambda||\mathbf{w}||_0$, where
$||\mathbf{w}||_0 = |\{i : w_i \ne 0\}|$; "Sparsity trick": use $||\mathbf{w}||_1$
Lasso:min $\frac{1}{n}\sum_{i=1}^n (y_i - \mathbf{w}^T\mathbf{x}_i)^2 + \lambda||\mathbf{w}||_1$(inclds FS)
L1-SVM: $\min_{\mathbf{w}} \frac{1}{n}\sum l_H(\mathbf{w}; \mathbf{x}_i, y_i) + \lambda||\mathbf{w}||_1$
Greedy: +any method, -slower (train many models);
L0/L1-regul.: +faster, -only lin models

Reprsnt.r Thm: $\hat{\mathbf{w}} = \sum_i \alpha_i(y_i)\mathbf{x}_i \in <\mathbf{x}_1, .., \mathbf{x}_n>$
⇒Perc.: $\min_\alpha \sum_i max(0, -y_i\sum_j \alpha_j y_j(\mathbf{x}_j^T\mathbf{x}_i))$
KT: 1. Use $\mathbf{w}$ in Thm as ansatz, replacing $\mathbf{w}$ with α; 2. $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T\mathbf{x}' \mapsto \phi(\mathbf{x})^T\phi(\mathbf{x}') =: k_\phi(\mathbf{x}, \mathbf{x}')$
PA:1.$\alpha := 0$. 2.$t = 1, 2, ...$: a)Pick $(x_i, y_i) \sim D$
b) $\alpha_i := \alpha_i + \eta_t\max(0, -\text{sgn}(y_i\sum_j \alpha_j y_j k(\mathbf{x}_j, \mathbf{x}_i)))$
Def: k kernel iff K sym. & pos. semi-def. iff SP/IP
· Poly: $(\mathbf{x}^T\mathbf{x}+1)^d$, Gaussian/RBF: $e^{1||\mathbf{x}-\mathbf{x}'||_2^2/h^2)}$,
Laplacian: $e^{-||\mathbf{x}-\mathbf{x}'||_1/h}$
· k_1+k_2, k_1k_2, ck_1 for $c>0$ and $f(k_1)$ for f poly with pos. coeffs or exponential are also kernels
· $(k_i)_i^d$ kernels ⇒ $k(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^d k_j(x_j, x_j')$ kernel
· $k((x,y),(x',y')) := k_1(x,y)k_2(x',y')$ kernel
· $k((x,y),(x',y')) := k_1(x,y) + k_2(x',y')$ kernel
P: $\hat{\alpha} = \arg\min_\alpha \frac{1}{n}\sum_i \max(0, -y_i\alpha^T\mathbf{k}_i)$
SVM: $\hat{\alpha} = \arg\min_\alpha \frac{1}{n}\sum_i \max(0, 1 - y_i\alpha^T\mathbf{k}_i) + \lambda\alpha^T\mathbf{D}_y\mathbf{K}\mathbf{D}_y\alpha$, $\mathbf{k}_i = [y_1k(\mathbf{x}_i, \mathbf{x}_1), ..., k(\mathbf{x}_i, \mathbf{x}_n)]$
RR: $\hat{\alpha} = \arg\min_\alpha \frac{1}{n}||\alpha^T\mathbf{K} - y||_2^2 + \lambda\alpha^T\mathbf{K}\alpha =$
$(\mathbf{K} + n\lambda\mathbf{I})^{-1}\mathbf{y}$ (closed form sol) **Kernel reg. pred.:** $\hat{y} = \sum_{j=1}^n \alpha_j k(x_j, x)$
Kernel bin. cl. pred.: $\hat{y} = \text{sgn}\left(\sum \alpha_j y_j k(x_j, x)\right)$
k-NN: $\hat{y}(\mathbf{x}) = \text{sgn}(\sum y_i 1[\mathbf{x}_i\ kNN\ of\ \mathbf{x}])$ (k?CV!)
+No training necessary, -depends on all data/ineff.
k-P: +Optim. weights improve perf., +Some k capture "global trends", +Depends only on wrongly classified ex.s, -Training requires optimization
Sum: Can derive non-para. m.s from para. w. k's
Prob: Paramatric models "rigid", non-param. fail to extrapolate: **Sol:** (Semi-param. m.) Add. comb. of lin. & non-lin. kernels
· E.g. $k(\mathbf{x}, \mathbf{x}') = c_1\exp(-||\mathbf{x}-\mathbf{x}'||_2^2/h^2) + c_2\mathbf{x}^T\mathbf{x}'$
$\Rightarrow f(\mathbf{x}) = \sum\alpha_i k(\mathbf{x}_i, \mathbf{x}) = f_\alpha(\mathbf{x}) + \mathbf{w}_\alpha^T\mathbf{x}$

Downsmplng:+Smaller/faster,-Wasteful/info-loss;
Upsmplng:+Uses $\forall(x,y)$,-slow,-adds artificial info;
Cost-sens. loss: $l_{CS}(\mathbf{w}; \mathbf{x}, y) = c_y l(\mathbf{w}; \mathbf{x}, y)$,$c_y > 0$.
Alt: Thresh τ in sgn$(\mathbf{w}^T\mathbf{x} - \tau)$. Acc.=$\frac{TP+TN}{n}$;
Prec.=$\frac{TP}{TP+FP}$; $\mathbb{E}[TPR] = p = \mathbb{E}[TPR]$;
Rec.=TPR=$\frac{TP}{TP+FN}$; FPR=$\frac{FP}{TN+FP}$;
F1-Score=$\frac{2TP}{2TP+FP+FN} = \frac{2}{Prec^{-1}+TPR^{-1}}$
Thm: $A1 \ge A2$ $ROC = \frac{TPR}{FPR}$ iff $A1 \ge A2$ $\frac{Prec.}{Rec.}$
Multi lin. class.: $\hat{y} = \arg\max_j \mathbf{w}_j^T\mathbf{x}$, $||\mathbf{w}_j|| =^! 1$.
Alt (1v1): $\hat{y} = \arg\max_{i\le c} |\{j : 0 < \text{sgn}(\mathbf{w}_{ij}^T\mathbf{x})\}|$
Encode:$1 \mapsto [0, ..., 1]$,$2 \mapsto [0, ..., 1, 0]$,$c \mapsto [1, ..., 1, 1]$
·reduces c or $c(c-1)/2$ req. bin. clas.rs to $O(\log_2 c)$
MCSVM: $\nabla l = x(1 - 2\cdot 1[\neg(*) \wedge i = y])1[(*) > 0]$
$l_{MC-H}(\mathbf{W}; \mathbf{x}, y) = \max(0, 1 + \max_{j\in[c]\setminus y} \mathbf{w}_j^T - \mathbf{w}_y^T\mathbf{x})$
Idea: Instead of cust. feats min $\sum l(y_i; \sum w_j\phi_j(\mathbf{x}_i))$ learn feat param.s: $\min_{\mathbf{w},\theta}\sum l(y_i; \sum w_j\phi(\mathbf{x}_i, \theta_j))$
· $\phi(\mathbf{x}, \theta) = \varphi(\theta^T\mathbf{x})$; φ = act. fun. e.g. $Sigm(z) = \frac{1}{1+\exp(-z)}$,$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$,ReLU=$\max(z, 0)$
ANN:nest.d comp (var) lin f.s comp w (fxd) nonlins

Forward propagation/ANN prediction:
1.For $j \in$ Layer$_1$, set $v_j = x_j$
2.For each layer $l = 1 : L-1$
For $j\in$ Layer$_l$, set:$v_j = \varphi(\sum_{i\in \text{Layer}_{l-1}} w_{j,i}v_i)$
3.For $j \in$ Layer$_L$: $f_j = \sum_{i\in\text{Layer}_{L-1}} w_{j,i}v_i$
4.Predict $y_j = f_j$ / $y_i = \text{sgn}(f_j)$ / $y_j = \arg\max_i f_i$
Or Alt: 1.For $\mathbf{v}^{(0)} := \mathbf{x}$
2.For $l = 1 : L-1$: $\mathbf{v}^{(l)} := \varphi(\mathbf{W}^{(l)}\mathbf{v}^{(l-1)})$
3. $\mathbf{f} := \mathbf{W}^{(L)}\mathbf{v}^{(L-1)}$
4.Predict $\mathbf{y} = \mathbf{f}$ / $\mathbf{y} = \text{sgn}(\mathbf{f})$ / $\mathbf{y} = \arg\max_i \mathbf{f}$
Thm (UAT): Let σ be a contin. sigm. func.. Then $\{G(x) = \sum_{j=1}^N \alpha_j\sigma(y_j^Tx + \theta_j)\} \subset^{dense} C^0([0,1]^n)$.
Prob: $\mathbf{W}^* = \arg\min_{\mathbf{W}}\sum l(\mathbf{W}; \mathbf{y}_i, \mathbf{x}_i)$ not convex.
Multi-loss: $l(\mathbf{W}; \mathbf{y}, \mathbf{x}) = \sum l_i(\mathbf{W}, y_i, \mathbf{x})$
SGD for ANNs: 1. Initialize weights $\mathbf{W}$
2. For $t = 1, 2, ..$:
Pick $(\mathbf{x}, \mathbf{y}) \in D$ U-randomly
Take step: $\mathbf{W} := \mathbf{W} - \eta_t\nabla_{\mathbf{W}} l(\mathbf{W}; \mathbf{y}, \mathbf{x})$

Backpropagation:
1.For $j\in$ Layer$_{L+1}$:
a) Compute error signal $\delta_j = l_j'(f_j)$
b) For each unit $i \in$ Layer$_L$: $\frac{\partial}{\partial w_{j,i}} = \delta_j v_i$
2.For $l = L-1 : 1$ and $j \in$ Layer$_l$:
a) Error signal: $\delta_j = \varphi'(z_j)\sum_{i\in\text{Layer}_{l+1}} w_{i,j}\delta_i$
b) For $i \in$ Layer$_{l-1}$: $\frac{\partial}{\partial w_{j,i}} = \delta_j v_i$

Backpropagation (Matrix version):
1.For $j \in$ Layer$_{L+1}$:
a) Compute error $\delta^{(L)} = \mathbf{l}'(\mathbf{f}) := [l'(f_1), .., l'(f_p)]$
b) Gradient $\nabla_{\mathbf{W}^{(\mathbf{L})}} l(\mathbf{W}; \mathbf{y}, \mathbf{x}) = \delta^{(L)}\mathbf{v}^{(L-1)T}$
2.For $l = L-1 : 1$:
a) Error: $\delta^{(l)} = \varphi'(\mathbf{z}^{(l)}) \cdot_{pw} (\mathbf{W}^{(l+1)T}\delta^{(l+1)})$
b) Gradient $\nabla_{\mathbf{W}^{(l)}} l(\mathbf{W}; \mathbf{y}, \mathbf{x}) = \delta^{(l)}\mathbf{v}^{(l-1)T}$
Init.: Keep Var[W] cnst acr. layers, avoid exp/van ∇
Glorot (tanh): $w_{i,j} \sim \mathcal{N}(0, \frac{1}{n_{in}})/\mathcal{N}(0, \frac{2}{n_{in}+n_{out}})$

He (ReLU): $w_{i,j} \sim \mathcal{N}(0, \frac{2}{n_{in}})$; **LR:** start fixed/small then decrease, e.g. $\eta_t = \min(0.1, 100/t)$ or decreasing step function; **Momentum:** (Escape loc. min.) $\mathbf{W} := \mathbf{W} - m\cdot a - \eta_t\nabla_{\mathbf{W}} l(\mathbf{W}; \mathbf{y}, \mathbf{x})$
Regul.:*Early stop. (when Err.(D_{val}) ↑),*Train dropout unit p/test $\mathbf{w} := p\mathbf{w}$, *$L(\mathbf{W}) + \lambda||\mathbf{W}||_F^2$
Batch norm.:(mini-batch $\mathcal{B} = (x_i)_i^m$) Learn γ, β.
For each layer: $(\varphi(wx) = \varphi(w\text{BN}_{\gamma,\beta}(x)))$
a) Normalize: $\hat{x_i} = \frac{1}{m}\sum(x_i - \mu_{\mathcal{B}})^2$
b) Scale & shift: $y = \gamma\hat{x}_i + \beta =: \text{BN}_{\gamma,\beta}(x_i)$
CNNs: Apply m diff. $f\times f$ filters to an $n\times n$ im. yields an $m\times l\times l$ to get, s.t. $l = \frac{n+2\cdot\text{padding}-f}{\text{stride}} + 1$
Past:sigmoid/tanh(difbl),**Now:**ReLu(fast,stable ∇s)
Kernels:+Convex,+noise robust,±$O(D)$,-1 layer;
ANNs:+flexible,nonlin,+layers(abstr),-may params and choices,-noise sensitive
k-Means: Pick centers of k clusters $\hat{\mu} = \arg\min\hat{R}$, where $\hat{R}(\mu) = \hat{R}(\mu_1, .., \mu_k) = \sum_i \min_j ||\mathbf{x}_i - \mu_j||_2^2$.
¬conv.⇒NP-h.But:Lloyd's (local) heuristic $O(knd)$:
1. Init. cluster centers: $\mu^{(0)} = [\mu_1^{(0)}, .., \mu_k^{(0)}]$
2. While not converged:
a) For $\mathbf{x}_i \in D$: $z_i^{(t)} = \arg\min_j ||\mathbf{x}_i - \mu_j^{(t-1)}||_2^2$
b) Update center as mean of assigned data pts
$\mu_j^{(t)} = \frac{1}{n_j}\sum_{i:z^{(t)}=j}\mathbf{x}_i$, where $n_j = |\{i : z_i^{(t)} = j\}|$
kMs++ seeding:($\mathbb{E}[\hat{R}(\mu^{(0)})] \le O(\log k)\min_\mu \hat{R}(\mu)$)
1. Start w. rand. pt. $\mathbf{x}_{i_1}$ as centr $\mu_1^{(0)} = \mathbf{x}_{i_1}$,
2. For $j = 2 : k$: Pick i_j with prob.:
$\frac{1}{C}\cdot\min_{1\le l\le j-1} d(\mathbf{x}_{i_j}, \mu_l^{(0)})$ and set $\mu_j^{(0)} = x_{i_j}$.
MS: Regul.,heuristic qu.m.s (elbow),info. theo. basis
PCA(k=1):$\arg\min_{||\mathbf{w}||=1}\{\sum_{i=1}^n ||z_i\mathbf{w} - \mathbf{x}_i||_2^2, z_i^* = \mathbf{w}^T\mathbf{x}_i\} =$
$\arg\max_{||\mathbf{w}||=1}\sum(\mathbf{w}^T\mathbf{x}_i)^2 \underset{\hat{\mu}=0}{=} \arg\max_{||\mathbf{w}||=1}\mathbf{w}^T\Sigma\mathbf{w} = \mathbf{v}_1$ princ.
EV of $\Sigma = \sum_i^d \lambda_i\mathbf{v}_i\mathbf{v}_i^T, \lambda_i \ge \lambda_{i+1} \ge 0$
$(f : d \to k > 1)$: **Sol:** $\mathbf{z}_i = \mathbf{W}^T\mathbf{x}_i = f(\mathbf{x}_i)$, $\Sigma =$"
$\arg\min_{\mathbf{W}\in\mathbf{O}^{||}(d\times k), \mathbf{Z}\in\mathbb{R}^{k\times n}}\sum||\mathbf{W}\mathbf{z}_i - \mathbf{x}_i||_2^2$
$\mathbf{W} := (\mathbf{v}_1|\cdot|\mathbf{v}_d) \in \mathbb{R}^{d\times k}$orth≡ $\mathbf{W}^T\mathbf{W} = \mathbf{I} \ne \mathbf{W}\mathbf{W}^T$
SVD:$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ ⇒ 1st k p.c. are 1st k cols of $\mathbf{V}$
((Pf: $n\mathbf{\Sigma} = \mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{S}^T\mathbf{U}^T\mathbf{U}\mathbf{S}\mathbf{V}^T = \mathbf{V}\mathbf{S}^T\mathbf{S}\mathbf{V}^T$))
K-PCA($k=1$): $\arg\max_\alpha\{\alpha^T\mathbf{K}^T\mathbf{K}\alpha : \alpha^T\mathbf{K}\alpha = 1\}$
Sol: $\alpha^* = \frac{\mathbf{v}_1}{\sqrt{\lambda_1}}$, $\mathbf{K} = \sum\lambda_i\mathbf{v}_i\mathbf{v}_i^T$, $\lambda_1 \ge \cdots\lambda_d \ge 0$
$(k\ge1)$:$\alpha^{(i)} = \frac{\mathbf{v}_i}{\sqrt{\lambda_1}} \in \mathbb{R}^n$ for $1 \le i \le k$, $\mathbf{K} =$",
$f(\mathbf{x}) = \mathbf{z} = (z_i)_i^k = (\sum_j \alpha_j^{(i)}k(\mathbf{x}_j, \mathbf{x}))_i^k$
Center: $\mathbf{K}' = \mathbf{K} - \mathbf{KE} - \mathbf{EK} + \mathbf{EKE}$, $\mathbf{E} = \frac{1}{n}\mathbf{1}\cdot\mathbf{1}^T$
Autoenc.s: Learn Id_d: $f(\mathbf{x};\theta) = f_2(f_1(\mathbf{x};\theta_1);\theta_2)$, s.t. $f_1 : \mathbb{R}^d \to \mathbb{R}^k$. **NNA:** take hidden layer as $f_1(\mathbf{x})$ train $\min_{\mathbf{W}}\sum||\mathbf{x}_i - \mathbf{f}(\mathbf{x}_i;\mathbf{W})||_2^2$ via bckprop SGD
$\varphi = Id$ (φ act. func)⇒ f =PCA solution
Probmod: $(\mathbf{x}_i, y_i) \sim P(\mathbf{X}, Y)$, $h : \mathcal{X} \to \mathcal{Y}$, risk:
$R(h) = \mathbb{E}_{\mathbf{X},Y}[l(y; h(\mathbf{x}))]$; **Reg.:**
$R(h) = \int P(\mathbf{x}, y)l(y; h(\mathbf{x}))d\mathbf{x}dy$;
Class.:$R(h) = \mathbb{E}[1[Y \ne h(\mathbf{X})]]$;
$h^*(x) = \arg\min_{\hat{y}}\mathbb{E}_Y[1[Y \ne \hat{y}|\mathbf{X} = \mathbf{x}]] =$
$\arg\max_{\hat{y}} P(Y = \hat{y}|\mathbf{X} = \mathbf{x})$

E.g. **LSR:**
$R(h) = \mathbb{E}_{X,Y}[(Y-h(\mathbf{X}))^2] = \mathbb{E}[\min_h \mathbb{E}[(Y - h(\mathbf{X}))^2|\mathbf{X}=\mathbf{x}]] \overset{\frac{dl}{dy}=0}{=} \mathbb{E}[(\mathbb{E}[Y|\mathbf{X}=\mathbf{x}] - h(\mathbf{X}))^2]$, i.e. $h^*(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X}=\mathbf{x}]$ (Bayes' opt. pred. for l^2)
Practice: $\hat{y} = \hat{\mathbb{E}}[Y|\mathbf{X}] = \int y \hat{P}(Y|\mathbf{X}) dy$
MLE: $\theta^* = \arg\max_\theta \hat{P}(y_1,...,y_n|\mathbf{x}_1,...,\mathbf{x}_n,\theta) \overset{iid}{=} \arg\min -\sum \log \hat{P}(y_i, \mathbf{x}_i, \theta)$. Easy to show ...
Thm:$f(y|\mathbf{x}) = \mathcal{N}(h^*(\mathbf{x}),\sigma^2)(y) \iff h^* = \hat{h}$ =LSE
e.g. $y_i \sim \mathcal{N}(\mathbf{w}^T\mathbf{x}_i, \sigma^2) \Rightarrow \hat{\mathbf{w}} = \arg\min \sum(y_i - \mathbf{x}_i)^2$
BV-T.o.: Pred.Err.=Bias^2+Var+Noise=Exp.risk= $\mathbb{E}_D\mathbb{E}_{\mathbf{X},Y}[(Y-\hat{h}_D(\mathbf{X}))^2]$;**Noise:**$\mathbb{E}_{\mathbf{X},Y}[(Y-h^*(\mathbf{X}))^2]$;
Bias:$\beta = \mathbb{E}_X[\mathbb{E}_D\hat{h}_D(\mathbf{X}) - h^*(\mathbf{X})]$; **Variance:**
$\mathbb{E}_{\mathbf{X}} \mathrm{Var}_D[\hat{h}_D(\mathbf{X})]^2 = \mathbb{E}_{\mathbf{X}}\mathbb{E}_D[\hat{h}_D(\mathbf{X}) - \mathbb{E}_{D'}\hat{h}_{D'}(\mathbf{X})]^2$; β(mle/lse)=0,use regul. trade bit of β for much Var.
MAP: $\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} P(\mathbf{w}|\mathbf{x}_{1:n}, y_{1:n}) \overset{Bayes'}{=} \arg\max_{\mathbf{w}} \frac{P(\mathbf{w})P(y_{1:n}|\mathbf{x}_{1:n},\mathbf{w})}{P(y_1,..,y_n|\mathbf{x}_1,...,\mathbf{x}_n)} = \arg\max_{\mathbf{w}} \log(\cdots)$
Thm:$\arg\min_{\mathbf{w}} \sum l(\mathbf{w}^T\mathbf{x}_i;\mathbf{x}_i,y_i) + C(\mathbf{w}) = \arg\max_{\mathbf{w}} \Pi_i P(y_i|\mathbf{x}_i,\mathbf{w})P(\mathbf{w}) = \arg\max_{\mathbf{w}} P(\mathbf{w}|D)$, $C(\mathbf{w}) = -\log P(\mathbf{w}), l(\mathbf{w}^T\mathbf{x}_i;y_i) = -\log P(y_i|\mathbf{x}_i,\mathbf{w})$
·MAP w. $(y|\mathbf{x},\mathbf{w}) \sim \mathcal{N}(\mathbf{w}^T\mathbf{x},\sigma^2) + w_i \sim \mathcal{N}(0,\beta^2)$ is $\hat{w} = \arg\min_{\mathbf{w}} \frac{\sigma^2}{\beta^2}||\mathbf{w}||_2^2 + \sum_i^n (y_i - \mathbf{w}^T\mathbf{x}_i)^2$ =RR$(\frac{\sigma^2}{\beta^2})$
· " w.Lapl.prior$(y|\mathbf{x},\mathbf{w}) \sim \mathcal{L}(\mathbf{w}^T\mathbf{x}, b) \Rightarrow \mathbf{w}$ =L1-RR
· Now come up w. new methods (improving robustness) by changing prior/l.f. e.g. Student's s-t
Log. reg.: $(y|\mathbf{x},\mathbf{w}) \sim \mathrm{Ber}(\sigma(\mathbf{w}^T\mathbf{x}))$ (i.e. Bernoulli noise) with $\sigma(z) := \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x})}$. Now using MLE, we find: $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum \log(1+\exp(-y_i\mathbf{w}^T\mathbf{x}_i)) = \arg\min_{\mathbf{w}} \sum l_{logistic}(\mathbf{w};\mathbf{x}_i,y_i) = \arg\min_{\mathbf{w}} \hat{R}(\mathbf{w})$
Logistic SGD: ($l_{\log} \approx l_P$ is convex & smooth! :-))
1. Init w:
2. For t=1,2,...:
   a) Pick (x,y) U-randomly
   b) Compute prob of misclass. with current model $\hat{P}(Y=-y|\mathbf{w},\mathbf{x}) = (1+\exp(y\mathbf{w}^T\mathbf{x}))$
   c) Take step: $\mathbf{w} = \mathbf{w} + \eta_t y \mathbf{x} \hat{P}(Y=-y|\mathbf{w},\mathbf{x})$

Again, use MAP to get, regularized methods:
l2 (Gauss. prior): $\min_{\mathbf{w}} \sum l_{\log}(\mathbf{w};y_i,\mathbf{x}_i)) + \lambda||\mathbf{w}||_2^2$
· c) Step:$\mathbf{w} = \mathbf{w}(1-2\lambda\eta_t) + \eta_t y\mathbf{x}\hat{P}(Y=-y|\mathbf{w},\mathbf{x})$
l1 (Laplace prior): $\min_{\mathbf{w}} \sum l_{\log}(\mathbf{w};y_i,\mathbf{x}_i)) + \lambda||\mathbf{w}||_1$
Classify: $\hat{P}(y|\mathbf{x},\hat{\mathbf{w}}) = (1+\exp(-y\hat{\mathbf{w}}^T\mathbf{x})$
K-LogR:$\hat{\alpha} = \arg\min_\alpha \sum \log\big(1 + \exp(-y_i\alpha^T\mathbf{K}_i)\big) + \lambda\alpha^T\mathbf{K}\alpha, \mathbf{K} = (\mathbf{K}_1|\cdots|\mathbf{K}_n)$
Classify: $\hat{P}(y|\mathbf{x},\hat{\alpha}) = \big(1+\exp(-y\sum\alpha k(\mathbf{x}_j,\mathbf{x})\big)^{-1}$
MC LogR:$P(Y=i|\mathbf{x},\mathbf{w}_{1:c}) = \frac{\exp(\mathbf{w}_i^T\mathbf{x})}{\sum\exp(\mathbf{w}^T\mathbf{x})}$ (w.l.o.g. $\mathbf{w}=0$ for uniqueness). **Cross-entropy loss:**
$l_{CE}(y;\mathbf{x},\mathbf{w}_{1:c}) = -\log P(Y=y|\mathbf{x},\mathbf{w}_1,...,\mathbf{w}_c)$
In ANN: $l_{CE}(Y=i,f_1,...,f_c) = -\log\frac{\exp(f_i)}{\sum_j^c\exp(f_j)}$
SVM/Perc:+sometimes higher accuracy,+sparse sol's;-MC class difficult; **LogR:**+Class probs;-Dense sols
Decision Theory: $C: \mathcal{Y}\times\mathcal{A} \to \mathbb{R}$, $\mathcal{A}$=Action set
Bayesian D.T.: Do $a^* = \arg\min_{a\in\mathcal{A}} \mathbb{E}_y[C(y,a)|\mathbf{x}]$
Bayesian optm. dec.: dec. taken if $P(y|\mathbf{x})$ known

E.g.s: Ass. $\hat{P}(y|\mathbf{x}) = \mathrm{Ber}(y;\sigma(\hat{\mathbf{w}}^T\mathbf{x}))$, $\mathcal{A}=\{\pm1\}$, $C(y,a) = 1[y\neq a] \Rightarrow a^* = \arg\min_{a\in\mathcal{A}} \mathbb{E}[C(y,a)|\mathbf{x}] \overset{\cdots}{=} \mathrm{sgn}(\mathbf{w}^T\mathbf{x}) = \arg\max_y \hat{P}(y|\mathbf{x})$ **Asym. Cs:** $\hat{P}, \mathcal{A}$=", $C(y,a) = c_{FP}1[y=-1,a=1] + c_{FN}1[y=1,a=-1]$ pred.1 $\iff c_+ = \mathbb{E}[C(y,1)|\mathbf{x}] < c_- = \mathbb{E}[C(y,-1)|\mathbf{x}] \iff p := P(y=+1|\mathbf{x}) > \frac{c_{FP}}{c_{FP}+c_{FN}}$
"Doubtful" LR: $\hat{P}$=", $\mathcal{A}=\{\pm1, D\}$, $C(y,a) = 1[y\neq a, a\in\{\pm1\}] + c1[a=D] \Rightarrow a^* = \arg\min_{a\in\mathcal{A}} \mathbb{E}_y[C(y,a)|\mathbf{x}] = y1_A + D1_{A^c}$, $A = [\hat{P}(y|\mathbf{x}) \geq 1-c]$,i.e.only pick likley class if sure
LS:$\hat{P}(y|\mathbf{x}) = \mathcal{N}(y;\hat{\mathbf{w}}^T\mathbf{x},\sigma^2), \mathcal{A}=\mathbb{R}, C(y,a) = (y-a)^2 \Rightarrow a^* = \arg\min_{a\in\mathcal{A}}\mathbb{E}_y[C(y,a)|\mathbf{x}] \overset{\partial=0,\int}{=} \hat{\mathbf{w}}^T\mathbf{x}$,
Asym. Cs: $\hat{P},\mathcal{A}$=", $C(y,a) =$ $c_1\max(y-a,0) + c_2\max(a-y,0)$ =underest+overest $\Rightarrow a^* = \arg\min_{a\in\mathcal{A}}\mathbb{E}_y[C(y,a)|\mathbf{x}] =$ $\hat{\mathbf{w}}^T\mathbf{x} + \sigma\Phi^{-1}(\frac{c_1}{c_1+c_2})$
Uncert. sampling: Pick ex. we r most uncert. bout, maintain uncert. $D_X = \{\mathbf{x}_1,...,\mathbf{x}_n\}$
`Init` $D=\emptyset$ |viol. iid, can get stuck w bad models
`For t=1,2,3...` | improve: red uncert of D
  `Est.` $\hat{P}(Y_i|\mathbf{x}_i)$ `given` D
  `Pick most uncer.:` $i_t = \arg\min_i |0.5 - \hat{P}(Y_i|\mathbf{x}_i)|$
  `Query label` y_{i_t} `and set` $D := D\cup\{(\mathbf{x}_{i_t}, y_{i_t})\}$
Gen. models: predict $P(\mathbf{x},y)$ instead of $P(y|\mathbf{x})$
1.Estimate prior on labels $P(y)$
2.Estimate $P(\mathbf{x}|y)\ \forall y \leq c (\Rightarrow P(\mathbf{x},y) = P(\mathbf{x}|y)P(y))$
3.Obtain $P(y|\mathbf{x}) = \frac{P(y)P(\mathbf{x}|y)}{Z}, Z = P(x)$ (Bayes)
Naive Bayes:$O(cd), \sum_y^c P(Y=y) = \sum_{y\in\mathcal{Y}} p_y = 1$
· Model feats: $P(X_1,..,X_d|Y) \overset{iid}{=} \prod_{j=1}^d P(X_j|Y)$, with $P(x_j|y) = \mathcal{N}(x_j|\mu_{y,j},\sigma^2_{y,j}), j\leq d$ (indep feats)
· Determine $P(Y), \mu_{j,y}, \sigma_{j,y}$ by MLE (given D)
· MLE 4 class prior: $\hat{p}_y = \frac{|\{i:y_i=y\}|}{n} =: \frac{n_y}{n}$
· MLE 4 feat distr.: $\hat{\mu}_{y,j} = \frac{1}{n_y}\sum_{i:y_i=y} x_{i,j}, (\mathbf{x}_i,y_i)$
$\hat{\sigma}^2_{y,j} = \frac{1}{n_y}\sum_{i:y_i=y}(x_{i,j}-\hat{\mu}_{y,j})^2$ · Prediction: $y = \arg\max_{y'}\hat{P}(y'|\mathbf{x}) = \arg\max_{y'}\hat{P}(y')\prod_j^d\hat{P}(x_j|y')$
·Bin. cl./$c=2 \Rightarrow y = \mathrm{sgn}\log\frac{P(Y=1|\mathbf{x})}{P(Y=-1|\mathbf{x})} =: \mathrm{sgn}\, f(\mathbf{x})$
· $c=2, p_1=p_2=.5, \sigma_{i,y} = \sigma_i \Rightarrow f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}+w_0$, $w_0 = \log\frac{\hat{p}_+}{1-\hat{p}_+} + \sum_j^d\frac{\hat{\mu}^2_{-,j}-\hat{\mu}^2_{+,j}}{2\hat{\sigma}^2_j}, w_j = \frac{\hat{\mu}_{+,j}-\hat{\mu}_{-,j}}{\hat{\sigma}^2_j}$
Prob: overconfidence, so dont use (contin) probs
Gn.GB:$O(cd^2), P(\mathbf{x}|y) \sim \mathcal{N}(\mu_y,\Sigma_y)$(NB:$\Sigma_y$ =diag)
· MLE 4 distr. $\hat{P}(\mathbf{x}|y) = \mathcal{N}(\mathbf{x};\hat{\mu}_y,\hat{\Sigma}_y)$ with $\hat{\Sigma}_y = \frac{1}{n_y}\sum_{i:y_i=y}(\mathbf{x}_i-\hat{\mu}_y)(\mathbf{x}_i-\hat{\mu}_y)^T$ and
· MLE 4 $\hat{p}_y$ and $\hat{\mu}_y = (\hat{\mu}_{y,j})_j^d$ same as 4 NB
· Prediction: same as with NB
·$c=2 \Rightarrow f(\mathbf{x}) = \log\frac{p}{1-p} + \frac{1}{2}[\log\frac{|\hat{\Sigma}_-|}{|\hat{\Sigma}_+|} + (\mathbf{x}-\hat{\mu}_-)^T\hat{\Sigma}_-^{-1}(\mathbf{x}-\hat{\mu}_-) - (\mathbf{x}-\hat{\mu}_+)^T\hat{\Sigma}_+^{-1}(\mathbf{x}-\hat{\mu}_+)]$
F's LDA$O(d)$:GB w. $c=2, p=.5, \hat{\Sigma}_- = \hat{\Sigma}_+ =: \hat{\Sigma}$
· Predict $y = \mathrm{sgn}(f(\mathbf{x})) = \mathrm{sgn}(\mathbf{w}^T\mathbf{x}+w_0)$ with $\mathbf{w} = \hat{\Sigma}^{-1}(\hat{\mu}_+ - \hat{\mu}_-), w_0 = \frac{1}{2}(\hat{\mu}_-^T\hat{\Sigma}^{-1}\hat{\mu}_- - \hat{\mu}_+^T\hat{\Sigma}^{-1}\hat{\mu}_+)$
$f(\mathbf{x}) = \log\frac{p(\mathbf{x})}{1-p(\mathbf{x})} \equiv p(x) = \sigma(f(\mathbf{x}))$ so this is LR!
vs LR:+outliers/$P(x)$, -norm. **X** else not robust

vs PCA: LDA proj. 1-d subsp. 2 max. $\frac{\mathrm{Var(between\ class)}}{\mathrm{Var(within\ class)}}$, PCA(k=1) only max. (all) var.
Categ.NB:$P(X_j=x|Y=y) = \theta^{(j)}_{x|y}, \sum_x\theta^{(j)}_{x|y} = 1$
· MLE 4 cl. label dstr. $\hat{P}(Y=y) = \hat{p}_y = \frac{n_y}{n}$
· MLE 4 dstr.feat.s $\theta^{(j)}_{x|y} = \frac{|\{X_j=x,Y=y\}|}{n_y}, O(exp(d))$!
·$y = \arg\max_{y'}\hat{P}(y'|\mathbf{x}) = \arg\max_{y'}\hat{P}(y')\prod\hat{P}(x_j|y)$
Mixed distr: NB doesnt require feat.s have i.d.;e.g: $P(x_{1:20}|y) = \prod_j^{10}\mathrm{Cat}(x_j|y,\theta)\prod_j^{10}\mathcal{N}(x_j;\mu_{j|y},\sigma^2_{j|y})$
Prior over param.s:
$P(\theta|y_{1:n}) = \frac{1}{Z}P(\theta)P(y_{1:n}|\theta), Z = \int P(\theta)P(y_{1:n})d\theta$
·Prior dstr. and l.h.f. **conjugate** if post. dstr. stays as prior; e.g. l.h.f.:Bin., Prior:$\beta(\theta;\alpha_+,\alpha_-)$, Obs: D_{n_+,n_-}, Post:$\beta(\theta;\alpha_++n_+,\alpha_-+n_-)$; **MAP:** $\hat{\theta} = \arg\max_\theta P(\theta|y_1,...,y_n;\alpha_+,\alpha_-) = \frac{\alpha_++n_+-1}{\alpha_++n_++\alpha_--n_--2}$; more e.g. ($\beta$,Ber),(Dir,Cat/MultiNom),(G.s,G.s);Use pairs as regul.s
GMMs:$P(\mathbf{x}|\mu,\Sigma,\mathbf{w}) = \sum_i^c w_i\mathcal{N}(\mathbf{x};\mu_i,\Sigma_i)$ (convex) $L(w_{1:k},\mu_{1:k},\Sigma_{w:k}) = -\sum_i^n\log\sum_j^k w_j\mathcal{N}(\mathbf{x}_i|\mu_j,\Sigma_j)$
Prob: Non-convex, Σ_j's must stay sym-pos-def, unlike GBCs z in $P(z,\mathbf{x}) = w_z\mathcal{N}(\mathbf{x}|\mu_z,\Sigma_z)$ is unobs.
Hard EM: `* Init params` θ, `* For t=1,2...:`
`E: Pred.cl.:` $\forall i: z_i^{(t)} = \arg\max_z P(z|\mathbf{x}_i,\theta^{(t-1)}) = \arg\max_z P(z|\theta^{(t-1)})P(\mathbf{x}_i|z,\theta^{(t-1)})$,
`M: Compute MLE as for GBC:`
$\theta^{(t)} = \arg\max_\theta P(D^{(t)}|\theta)$ w. $D = ((\mathbf{x}_i, z_i^{(t)}))_i$.
PP:$\gamma_j(\mathbf{x}) := P(Z=j|\mathbf{x},\Sigma,\mu,\mathbf{w}) \overset{B's}{=} \frac{w_jP(\mathbf{x}|\Sigma_j,\mu_j)}{\sum_l w_lP(\mathbf{x}|\Sigma_l,\mu_l)}$
$(\mu^*,\Sigma^*,w^*) = \arg\min -\sum_i^n\log\sum_j^k w_j\mathcal{N}(\mathbf{x}_i|\mu_j,\sigma_j)$
s.t. $\mu_j^* = \frac{\sum\gamma_j(\mathbf{x}_i)\mathbf{x}_i}{\sum\gamma_j(\mathbf{x}_i)}$,
$\Sigma_j^* = \frac{\sum\gamma_j(\mathbf{x}_i)(\mathbf{x}_i-\mu_j^*)(\mathbf{x}_i-\mu_j^*)^T}{\sum\gamma_j(\mathbf{x}_i)}, w_j^* = \frac{1}{n}\sum\gamma_j(\mathbf{x}_i)$
(For SSL add. require $\gamma_j(\mathbf{x}_i) = 1[j=y_i]$ if y known)
Soft-EM:`While not converged:` |E=exp.suff.st.
`E:Calculate` $\gamma_j^{(t)}(\mathbf{x}_i)$ `given` $\mu^{(t)},\Sigma^{(t)},\mathbf{w}^{(t)}$(SSL:+y)
$\gamma_j^{(t)}(\mathbf{x}_i) = \frac{w_jP(\mathbf{x}_i|\Sigma_j,\mu_j)}{\sum w_lP(\mathbf{x}_i|\Sigma_l,\mu_l)}(/1\{j=y_i\}$ if y_i known)
`M:Fit clusters to weighted data` |M=Max.l.h.sol
$w_j^{(t)} = \frac{1}{n}\sum\gamma_j^{(t)}(\mathbf{x}_i), \mu_j^{(t)} = \frac{\sum\gamma_j^{(t)}(\mathbf{x}_i)\mathbf{x}_i}{\sum\gamma_j^{(t)}(\mathbf{x}_i)}$,
$\Sigma_j^{(t)} = \frac{\sum\gamma_j^{(t)}(\mathbf{x}_i)(\mathbf{x}_i-\mu_j^{(t)})(\mathbf{x}_i-\mu_j^{(t)})^T}{\sum\gamma_j^{(t)}(\mathbf{x}_i)}$(Optn.add.on$\Sigma$)
We can avoid degenerecy/ovrftng by $\Sigma_j^{(t)} += \nu^2\mathbf{I}$
Initing?$w^{(0)}$ ∼U,$\mu^{(0)}$ =rand/k-M++,σ =sph./$\hat{\sigma}^2 I$
(CV 4 k works i.C.2 kMs, try max log-l.h. on D_{val})
Soft(asgn)EM: higher l.h.s cuz better w cltr ovrlps
$\lim_{\sigma\to0}$SEM(σ^2I) ∼k-M;L's h∼HEM(sph.$\Sigma=\sigma^2I$)
Alg Props: Mon.incr. l.h.;GMM guar. 2 conv. loc.
GMBC: $P(\mathbf{x}|y) = \sum_{j=1}^{n_y} w_j^{(y)}\mathcal{N}(\mathbf{x};\mu_j^{(y)},\Sigma_j^{(y)})$
Dens.est.: Model $P(\mathbf{x})$ w. GMM but $P(y|\mathbf{x})$ discr.ly
Thm: EM equiv to following procedure:
·E-step: $Q(\Theta;\Theta^{(t-1)}) =$
$\mathbb{E}_{z_{1:n}}[\log P(x_{1:n},z_{1:n}|\Theta)|x_{1:n},\Theta^{(t-1)}]$
$= \sum_i^n\sum_{z_i}^k\gamma_{z_i}(x_i)\log P(x_i,z_i|\Theta)$
·M-step: $\Theta^{(t)} = \arg\max_\theta Q(\Theta;\Theta^{(t-1)})$

Thm: EM monot.ly incr. l.h. ($=P(x_{1:n},z_{1:n}|\Theta)$).
Cor: For Gauss mixt. this means EM guar. loc. conv.
Impl. gen. mod.s: Learn $\mathbf{X} = G(\mathbf{Z};\mathbf{w})$ w. Z "simple" distr data (e.g. $P(Z) = \mathcal{N}(0,\Sigma)$), G flexible nonlin func (e.g. NN) **GANs:** Train G w. noise Z while training D discriminator 2 distg. between X from G and raw real X
"Use discriminative learning to train generative model!"
$D: \mathbb{R}^d \to [0,1]$ wants: $D(x) = 1[x$ is real];
$G: \mathbb{R}^m \to \mathbb{R}^d$ wants: $D(G(z)) = 1$
Objective: $\min_{\mathbf{w}_G}\max_{\mathbf{w}_D} M(\mathbf{w}_G,\mathbf{w}_D) = \min_{\mathbf{w}_G}\max_{\mathbf{w}_D}\mathbb{E}_{\mathbf{x}\sim\mathrm{Data}}\log D(\mathbf{x};\mathbf{w}_D) + \mathbb{E}_{\mathbf{z}\sim\mathcal{N}}\log(1-D(G(\mathbf{z};\mathbf{w}_G);\mathbf{w}_D))$ (Saddle pt./GT)
Simul. (mini batch) GD:
$\mathbf{w}_G^{(t+1)} = \mathbf{w}_G^{(t)} - \eta_t\nabla_{\mathbf{w}_G}M(\mathbf{w}_G,\mathbf{w}_D^{(t)})$,
$\mathbf{w}_D^{(t+1)} = \mathbf{w}_D^{(t)} + \eta_t\nabla_{\mathbf{w}_D}M(\mathbf{w}_G^{(t)or(t+1)},\mathbf{w}_D)$
Probs: Data mem.⇒degen. sols, oscillations/divergence, mode collapse (special. on feat.s), "cannot compute l.h. on holdout set
.
PSt:$X_i \sim f(\cdot|\theta_0)$ w. unknown $\theta_0\in\Theta$ the likelihood func. (l.h.(f.)) is $L(\theta) = f(x_1,..,x_n|\theta) \overset{iid}{=} \prod_i^n f(x_i|\theta)$
The MLE is then $\hat{\theta} = \arg\max_{\theta\in\Theta}$
·$\varphi(\cdot\cdot) \leq t\varphi(x_1) + (1-t)\varphi(x_2) \Rightarrow \varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$
·$k(x,y) := (x^Ty)^m, x,y\in\mathbb{R}^d$ impl. repr. mon.s of degree $=m$, there are $\binom{d+m-1}{d}$, $O(d^m)\to O(d)$
·$k(x,y) := (1+x^Ty)^m$ impl. repr. mon.s of degree $\leq m$ there are $\binom{d+m}{m}$
Eigen.: $A\in\mathbb{R}^{n\times n}$, ∃EV basis $\Rightarrow A = QDQ^{-1}$
LU: Gauss. Elim., $A\in\mathbb{R}^{m\times n} \Rightarrow A = LU$
QR/QU: Gr.-Schm., $A\in\mathbb{R}^{n\times n} \Rightarrow A = QU$
SVD:$A\in\mathbb{R}^{m\times n}, U\in\mathbb{R}^{m\times m}, V\in\mathbb{R}^{n\times n}, \Sigma\in\mathbb{R}^{m\times n}: A = U\Sigma V^T$
Cholesky: $A\in\mathbb{R}^{n\times n}$, sym-pos-def $\Rightarrow A = LL^T$
Pos-def: $\forall x\in\mathbb{R}^n\setminus0: x^TAx > 0 \iff \sigma(A)\subset\mathbb{R}_{>0}$
·$\exists!x: Ax = b \iff \det A\neq0 \iff [Ax = 0 \equiv x = 0$
·$\det(\mathbf{X}^T\mathbf{X})\neq0 \iff \exists!\mathbf{w} = \arg\min\sum(y_i-\mathbf{w}^T\mathbf{x}_i)^2$
Counter-example: $\mathcal{D} = \{(0,0)\}$ has ∞ sols
·$\frac{\partial\mathbf{x}^T\mathbf{a}}{\partial\mathbf{x}} = \mathbf{a}$; $\frac{\partial\mathbf{a}^T\mathbf{X}\mathbf{b}}{\partial\mathbf{X}} = \mathbf{a}\mathbf{b}^T$; $\frac{\partial\mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial\mathbf{x}} = (\mathbf{A}+\mathbf{A}^T)\mathbf{x}$
LTP: $\bigsqcup_{\mathbb{N}} A = \Omega \implies P(B) = \sum P(B|A_i)P(A_i)$
Bayes' rule: $P(A|B) = P(B|A)\frac{P(A)}{P(B)}$;
$P(A_1,..,A_n) = P(A_1)P(A_2|A_1)\cdots P(A_n|A_1,..,A_{n-1})$
Norm Distr.: $p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$;
$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{|\det(\Sigma)|}}\exp(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu))$
C.M.: $\Sigma = \mathbb{E}[(\mathbf{X}-\mu)(\mathbf{X}-\mu)^T] = (\mathrm{Cov}(X_i,X_j))_{ij}$
$\mu = \mathbb{E}[\mathbf{X}]$; $\mathrm{Var}(\mathbf{a}^T\mathbf{X}) = \mathbf{a}^T\Sigma\mathbf{a}, \mathbf{a}\in\mathbb{R}$
Beta:$f(x) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}x^{\alpha_1-1}(1-x)^{\alpha_2-1}1_{(0,1)}$,
$\Gamma(a)\int_0^\infty t^{a-1}e^{-1}dt$; **Lapl Distr:**$\frac{1}{2b}\exp(-\frac{|x-\mu|}{b})$
St.'s-t:$\Gamma(\frac{\nu+1}{2})/(\sqrt{\nu\pi}\Gamma(\frac{\nu}{2}))(1+\frac{x^2}{\nu})^{-\frac{\nu+1}{2}}$
MN(w. repl.):$\frac{N!}{n_1!\cdots n_k!}p_1^{n_1}\cdots p_k^{n_k}$
MvHG(wo. repl.):$(\prod_{i=1}^c\binom{K_i}{k_i})/\binom{N}{n}$