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 $\Rightarrow$ 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 $\Rightarrow\hat{R}(\mathbf{w})\xrightarrow{|D|\to\infty}R(\mathbf{w})$ 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 **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: $f(\mathbf{x})=f_{\mathbf{w}}(\mathbf{x})=\text{sgn}(\mathbf{w}^T\mathbf{x})$, $l=l_{0/1}(\mathbf{w};\mathbf{x}_i,y_i):=1[y_i\neq f_{\mathbf{w}}(\mathbf{x}_i)]$ (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 $\mathbf{w}_{t+1}=\mathbf{w}_t-\eta_t\nabla l(\mathbf{w}_t;\mathbf{x}',y')$
Convergence: Guaranteed if $\sum_t\eta_t=\infty$ and $\sum_t\eta_t^2<\infty$. E.g.: $\eta_t=\frac{1}{1+t}$ or $\min(c,\frac{c'}{1+t})$.

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 $||\mathbf{w}||_0=|\{i:w_i\neq 0\}|$; "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: $\hat{\mathbf{w}}=\sum_i\alpha_i(y_i)\mathbf{x}_i\in<\mathbf{x}_1,..,\mathbf{x}_n>$
$\Rightarrow$**Perc.**: $\min_\alpha\sum_i max(0,-y_i\sum_j\alpha_j y_j(\mathbf{x}_j^T\mathbf{x}_i))$
Idea: 1. Use $\mathbf{w}$ in Thm as ansatz, replacing $\mathbf{w}$ with α; 2. $k(\mathbf{x},\mathbf{x}')=\mathbf{x}^T\mathbf{x}'\mapsto\phi(\mathbf{x})^T\phi(\mathbf{x}')=:k_\phi(\mathbf{x},\mathbf{x}')$
SHOULD I ADD KERNELIZED PERCEPTRON/MERCER/POS DEF DEF n CHAR n DECOMP (L7) HERE?
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 $\Rightarrow k(\mathbf{x},\mathbf{x}')=\sum_{j=1}^d k_j(x_j,x_j')$ 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
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
SHOULD I ADD KERNELIZED SVM & LR (L8/9) HERE??
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: Vary threshold τ in sgn$(\mathbf{w}^T\mathbf{x}-\tau)$
Acc.=$\frac{TP+TN}{n}$; Prec.=$\frac{TP}{TP+FP}$; Rec.=$\frac{TP}{TP+FN}$;
F1-Score=$\frac{2TP}{2TP+FP+FN}=\frac{2}{Precision^{-1}+Recall^{-1}}$;
TPR=$\frac{TP}{TP+FN}$; FPR=$\frac{FP}{TN+FP}$; $\to\mathbb{E}[\cdot PR]=p$
Thm: $A1\ge A2$ $ROC=\frac{TPR}{FPR}$ iff $A1\ge A2$ $\frac{Prec.}{Rec.}$
Multi lin. class.: $\hat{y}=\arg\max_j\mathbf{w}_j^T\mathbf{x}$, $||\mathbf{w}_j||\overset{!}{=}1$.
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 **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 $\nabla_{\mathbf{W}^{(L)}}l(\mathbf{W};\mathbf{y},\mathbf{x})=\delta^{(L)}\mathbf{v}^{(L-1)T}$
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.) $\mathbf{W}:=\mathbf{W}-m\cdot a-\eta_t\nabla_{\mathbf{W}}l(\mathbf{W};\mathbf{y},\mathbf{x})$
Overfit.: Early stopping (if ∂Err.$(D_{val})>0$), regul: $\min_{\mathbf{W}}\sum l(\mathbf{W};D_i)+\lambda||\mathbf{W}||_F^2$, Dropout $p(\text{unit},t)=\frac{1}{2}$
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: $y=\gamma\hat{x_i}+\beta=:\text{BN}_{\gamma,\beta}(x_i)$
CLs: 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,$\pm O(D)$,-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 $\hat{R}(\mu)=\hat{R}(\mu_1,..,\mu_k)=\sum_i\min_j||\mathbf{x}_i-\mu_j||_2^2$.
$\neg$conv.$\Rightarrow$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 $\mu_j^{(t)}=\frac{1}{n_j}\sum_{i:z_i^{(t)}=j}\mathbf{x}_i$, where $n_j=|\{i:z_i^{(t)}=j\}|$
kMs++:(Use2init:$\mathbb{E}[\hat{R}(\mu^{(0)})]=O(\log k)\min_\mu\hat{R}(\mu)$)
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.: $\frac{1}{C}\min_{1\le l\le j-1}d(\mathbf{x}_{i_j},\mu_l^{(0)})$ and set $\mu_j^{(0)}=x_{i_j}$.
MS:Regul.,heuristic qu.m.s (elbow),info. theo. basis
Lin dim red: $\arg\min_{\mathbf{z},||\mathbf{w}||=1}\sum_{i=1}^n||z_i\mathbf{w}-\mathbf{x}_i||_2^2$, $z_i^*=\mathbf{w}^T\mathbf{x}_i$
$\iff\arg\max_{||\mathbf{w}||=1}\sum(\mathbf{w}^T\mathbf{x}_i)^2\overset{\hat{\mu}=0}{\iff}\arg\max_{||\mathbf{w}||=1}\mathbf{w}^T\Sigma\mathbf{w}$,
where $\hat{\mu}=\frac{1}{n}\sum_i\mathbf{x}_i=\mathbb{E}[\mathbf{x}_i]\Rightarrow\Sigma=\frac{1}{n}\sum\mathbf{x}_i\mathbf{x}_i^T$
Sol: $\mathbf{w}^*=\mathbf{v}_1$ of $\Sigma=\sum_i^d\lambda_i\mathbf{v}_i\mathbf{v}_i^T,\lambda_{i+1}\ge\lambda_i\ge 0$
PCA($f:d\to k>1$): $\arg\min_{\mathbf{W},\mathbf{z}}\{\sum||\mathbf{W}\mathbf{z}_i-\mathbf{x}_i||_2^2$: $\mathbf{W}=(\mathbf{v}_1|\cdot|\mathbf{v}_d)\in\mathbb{R}^{d\times k}$orth$\}$**Sol**:$\mathbf{z}_i=\mathbf{W}^T\mathbf{x}_i=f(\mathbf{x}_i)$
SVD: $\mathbf{X}=\mathbf{U}\mathbf{S}\mathbf{V}^T$, where $\mathbf{U}\in\mathbb{R}^{n\times n}$ and $\mathbf{V}=(\mathbf{v}_1|\cdots|\mathbf{v}_d)\in\mathbb{R}^{d\times d}$ orth and $\mathbf{S}=diag(\lambda_1,...,\lambda_{\min(n,d)})\in\mathbb{R}^{n\times d}$
K-PCA($k=1$): $\arg\max_\alpha\{\alpha^T\mathbf{K}^T\mathbf{K}\alpha:\alpha^T\mathbf{K}\alpha=1\}$
Sol: $\alpha^*=\frac{\mathbf{v}_1}{\sqrt{\lambda_1}}$, $\mathbf{K}=\sum\lambda_i\mathbf{v}_i\mathbf{v}_i^T$, $\lambda_1\ge\cdots\lambda_d\ge 0$
$(k\ge 1)$:$\alpha^{(i)}=\frac{\mathbf{v}_i}{\sqrt{\lambda_1}}\in\mathbb{R}^n$ 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 K: $\mathbf{K}'=\mathbf{K}-\mathbf{KE}-\mathbf{EK}+\mathbf{EKE}$
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 $f_1(\mathbf{x})$ trning $\min_{\mathbf{W}}\sum||\mathbf{x}_i-\mathbf{f}(\mathbf{x}_i;\mathbf{W})||_2^2$ via bckprop SGD
$\varphi=Id\Rightarrow f=$PCA solution
Probmod: $(\mathbf{x}_i,y_i)\sim P(\mathbf{X},Y),h:\mathcal{X}\to\mathcal{Y}$, risk: $R(h)=\int P(\mathbf{x},y)l(y;h(\mathbf{x}))d\mathbf{x}dy=\mathbb{E}_{\mathbf{X},Y}[l(y;h(\mathbf{x}))]$
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. $y_i \sim \mathcal{N}(\mathbf{w}^T\mathbf{x}_i, \sigma^2) \Rightarrow \hat{\mathbf{w}} = \arg\min \sum(y_i - \mathbf{x}_i)^2$

Bias Variance Tradeoff:

Prediction error = Bias^2+Variance+Noise

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

$$E = mc^2.$$

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

$$\begin{aligned} a &= b \\ &= c \\ &= d. \end{aligned}$$

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

$$E = mc^2.$$

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

$$\begin{aligned} a &= b \\ &= c \\ &= d. \end{aligned}$$

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

$$E = mc^2.$$

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

$$\begin{aligned} a &= b \\ &= c \\ &= d. \end{aligned}$$

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

$$E = mc^2.$$

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

$$\begin{aligned} a &= b \\ &= c \\ &= d. \end{aligned}$$

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

$$E = mc^2.$$

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

$$\begin{aligned} a &= b \\ &= c \\ &= d. \end{aligned}$$

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

$$E = mc^2.$$

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

$$\begin{aligned} a &= b \\ &= c \\ &= d. \end{aligned}$$

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

$$E = mc^2.$$

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

$$\begin{aligned} a &= b \\ &= c \\ &= d. \end{aligned}$$

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

$$E = mc^2.$$

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

$$\begin{aligned} a &= b \\ &= c \\ &= d. \end{aligned}$$