Linear Regression

Model: $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$ ($+w_0$ with bias).
OLS: $L(\boldsymbol{w}) = \|X\boldsymbol{w} - \boldsymbol{y}\|^2$ ($X \in \mathbb{R}^{n\times d}$) is quadratic $\Rightarrow$ cvx $\Rightarrow$ any local min is global.
Normal eq. $X^\top X\boldsymbol{w} = X^\top \boldsymbol{y}$: always solvable, minimizer set $\mathcal{W} \neq \emptyset$; $|\mathcal{W}| \in \{1, \infty\}$.
Unique $\Leftrightarrow$ $X^\top X$ invertible $\Leftrightarrow$ rank $X = d$ (full col. rank: $n \geq d$ & no collinear cols; $n \geq d$ alone *not* enough): $\hat{\boldsymbol{w}} = (X^\top X)^{-1} X^\top \boldsymbol{y}$.
Else ($X^\top X$ singular): ∞ many $= \hat{\boldsymbol{w}} +$ null$(X^\top X)$. $n < d \implies X^\top X$ singular, and $\exists \boldsymbol{w}: X\boldsymbol{w} = \boldsymbol{y}$ (interpolates).
Adding features: train. loss never $\uparrow$.
1D ($w_0 + w_1 x$): $w_1^* = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)}$, $w_0^* = \bar{y} - w_1^* \bar{x}$.
Stat model: $\boldsymbol{y} = X\boldsymbol{w}^* + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I)$ $\Rightarrow \hat{\boldsymbol{w}} \sim \mathcal{N}(\boldsymbol{w}^*, \sigma^2 (X^\top X)^{-1})$ (unbiased).
Loss Functions: **Sqrd:** outliers; **Abs:** robust; not diff. at 0. **Huber:** ($\ell_H / \ell_{abs} \to \delta$); robust, smooth.

Optimization

GD (batch, all N): $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta \nabla f(\boldsymbol{w}^t)$; for smooth f, small enough $\eta > 0$ always $\downarrow f$. Exact but $O(N)$/step
SGD (1 example): unbiased est of ∇L, but noisy: can $\uparrow$ the loss, leave a stationary pt., be large even at an ε-stationary pt. of L. $O(1)$/step $\Rightarrow$ cheap; **Mini-batch SGD**: compromise; cost $\approx$ SGD if parallelized.
Momentum: ($\beta \in \mathbb{R}$): $\boldsymbol{w}^{t+1} - \boldsymbol{w}^t = \beta(\boldsymbol{w}^t - \boldsymbol{w}^{t-1}) - \eta \nabla_{\boldsymbol{w}} L(\boldsymbol{w}^t)$
Convexity (cvx): f cvx $\Leftrightarrow$ $D^2 f \succeq 0$ (PSD) $\forall \boldsymbol{w}$. If diff. & $\nabla f(\boldsymbol{w}) = 0 \Rightarrow \boldsymbol{w}$ global min. α-strong cvx $\Leftrightarrow$ $D^2 f \succeq \alpha I$ ($\alpha > 0$, PD) $\Rightarrow$ unique global min.
Preserved: (f, g cvx); $\alpha f + \beta g$ ($\alpha, \beta \geq 0$); $\max(f, g)$; affine compos. $f(A\boldsymbol{x} + \boldsymbol{b})$.
Not preserved: $\min(f, g)$; $f \cdot g$; $f(g(\boldsymbol{x}))$ (cvx iff g affine OR f cvx & non-decr).
Cvx any $\|\cdot\|$; ReLU; linear/affine. **Strict:** $\|\cdot\|^2$ (*strongly*); e^{-z}, $\log(1+e^{-z})$ (*not strongly*).

Model Evaluation & Selection

True model: f^*, $Y = f^* + \epsilon$, $\mathrm{Var}(\epsilon) = \sigma^2$
Gen. error: $L(\hat{f}; \mathbb{P}) = \mathbb{E}_{X,Y} \ell(\hat{f}(X), Y) = \mathbb{E}_X[(\hat{f}(X) - f^*(X))^2] + \sigma^2$ (estim. err. + noise.) BUT **Training err. underestimates gen. err:** use indep. test set: **test err.** $\to L(\hat{f}; \mathbb{P})$ (LLN) **Pitfall**: select a model by test err. $\Rightarrow$ test err. no longer unbiased est. of gen. err.
K-fold CV: $\mathrm{CV}_K(M) = \frac{1}{K}\sum_k L(\hat{f}_{M,k}; \mathcal{D}_k)$; pick $\arg\min_M \mathrm{CV}_K$; refit on all data, eval. on $\mathcal{D}_{\text{test}}$. $\uparrow K$: sub-models $\hat{f}_k \to$ full model, but each $\hat{L}_k$ less precise & **more** compute. LOOCV ($K = n$): max cost.
$\uparrow$val. set $\Rightarrow$ lower-variance gen-err estimate; $\uparrow$train set $\Rightarrow$ lower gen err. Train err $\leq$ val err usually, **not always**.
CV/LOOCV estimates the **method** M, not one fixed $\hat{f}$ (each fold a different f); deterministic if M is.

Regularization & Model Selection

$\downarrow$ model complexity $\Rightarrow \uparrow$ bias and gen. err.
$\uparrow$ complexity $\Rightarrow \downarrow$ train err (always), but $\uparrow$ variance and gen. err. $\Rightarrow$ bias-variance tradeoff
Bias-Var Decomp: $\mathbb{E}[\text{gen. err.}] = \underbrace{(\mathbb{E}[\hat{f}_{\mathcal{D}}] - f^*)^2}_{\text{bias}^2} + \underbrace{\mathbb{E}[(\hat{f}_{\mathcal{D}} - \mathbb{E}[\hat{f}_{\mathcal{D}}])^2]}_{\text{variance}} + \underbrace{\sigma^2}_{\text{noise}}$
Ridge & LASSO: $\min_{\boldsymbol{w}} \|\boldsymbol{y} - X\boldsymbol{w}\|^2 + \lambda \|\boldsymbol{w}\|$ $\equiv$ constrained $\min \|\boldsymbol{y} - X\boldsymbol{w}\|^2$ s.t. $\|\boldsymbol{w}\| \leq R$. $\uparrow \lambda \Leftrightarrow \downarrow R$.
$\uparrow \lambda$: $\uparrow$bias, $\downarrow$variance; $\widehat{\boldsymbol{w}} \to \boldsymbol{0}$ as $\lambda \to \infty$; $\lambda = 0$: OLS. Gen. err. non-monotonic, min at λ^*.
Ridge (L2), $\lambda \|\boldsymbol{w}\|_2^2$: $\widehat{\boldsymbol{w}}_\lambda = (X^\top X + \lambda I)^{-1} X^\top \boldsymbol{y}$ — closed form, always exists ($X^\top X + \lambda I \succ 0$), differentiable.
LASSO (L1), $\lambda \|\boldsymbol{w}\|_1$: **sparsity**. No closed form, not differentiable. Cvx relaxation of ℓ_0 (best-subset, infeasible).
Orthonormal X ($X^\top X = I$): soft-threshold $\hat{w}_j = \mathrm{sign}(w_j^{\mathrm{LS}}) \max(0, |w_j^{\mathrm{LS}}| - \lambda)$, $\boldsymbol{w}^{\mathrm{LS}} = X^\top \boldsymbol{y} \Rightarrow$ once $\hat{w}_j \neq 0$ at λ, stays $\neq 0$ for all $\lambda' < \lambda$.
Correlated features: L2 splits weight equally ($x_1 = x_2 \Rightarrow \hat{w}_1 = \hat{w}_2$, unique); L1 unstable: non-unique, only $\hat{w}_1 + \hat{w}_2$ fixed, drives one to 0.
Selecting Hyperparameters: Tune λ by CV over grid, ex. log-scale $\{10^{-3}, \ldots, 10^2\}$.

Classification

$\hat{y} = \mathrm{sign}\, f(\boldsymbol{x})$. $\ell_{0\text{-}1} = \mathbb{I}[\hat{y} \neq y] = \mathbb{I}[z < 0]$ for $z = yf(\boldsymbol{x}) = y\langle \boldsymbol{w}, \boldsymbol{x} \rangle$ ($= |\langle \boldsymbol{w}, \boldsymbol{x} \rangle|$ if correct; $\forall i$ iff lin. sep.). **Decision boundary** $\{\boldsymbol{x} : f(\boldsymbol{x}) = 0\}$; $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \phi(\boldsymbol{x}) \rangle$: bound. linear in $\phi(\boldsymbol{x})$, not $\boldsymbol{x}$.
Surrogate Losses: 0-1 loss non-cvx & non-diff. $\Rightarrow$ train on cvx, diff., $\downarrow$-in-z S.L. (eval still 0-1). **Logistic**: $\log(1+e^{-z})$: bounded grad, stable. **Exp**: e^{-z}: grad explodes as $z \to -\infty$, outlier-sensitive. **Logistic regression**: $\min_{\boldsymbol{w}} \frac{1}{n}\sum_i \log(1+e^{-z_i})$ (cvx). Sep. data: no finite min, $\|\boldsymbol{w}\| \to \infty$, but GD $\to \boldsymbol{w}_{MM}$ in **direction**; $+$L2 $\Rightarrow$ finite min. Non-sep.:unique glob min, GD converges to it.
Max-margin ($\boldsymbol{w}_{MM}$) & SVM: Solvable iff lin. sep.: $\mathrm{marg}(\boldsymbol{w}) = \min_i z_i$ st. $\|\boldsymbol{w}\|_2 = 1$; dist. to bound. $= \frac{|\langle \boldsymbol{w}, \boldsymbol{x} \rangle|}{\|\boldsymbol{w}\|_2}$. $\boldsymbol{w}_{MM} = \arg\max_{\|\boldsymbol{w}\|_2 = 1} \min_i z_i$. **Hard-margin SVM** $\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2$ s.t. $z_i \geq 1\ \forall i$, unique ($\|\cdot\|^2$ strictly cvx). Support vectors: $z_i = 1$; dist. to bound. $= 1/\|\boldsymbol{w}_{SVM}\|$. $\boldsymbol{w}_{SVM}/\|\boldsymbol{w}_{SVM}\| = \boldsymbol{w}_{MM}$.
Always solvable: **Soft-margin** $\min \|\boldsymbol{w}\|^2 + \lambda \sum_i \max(0, 1 - z_i)$ (ℓ_2 + **hinge**), unique $\boldsymbol{w}$. $\lambda \to \infty$: hard-marg; $\lambda \to 0$: $\boldsymbol{w} \to 0$.
Multiclass: K models, $\hat{y} = \arg\max_k f_k(\boldsymbol{x})$. **One-vs-rest**: K binary fits. **Cross-entropy** (joint): $\hat{p}_k = \mathrm{softmax}(f)_k = \frac{e^{f_k(\boldsymbol{x})}}{\sum_j e^{f_j(\boldsymbol{x})}}$ (k: class index); $\ell_{ce} = -\log \hat{p}_y = -\sum_k y_k \log \hat{p}_k$ (y one-hot, $y_k = 1$ iff k true class); $=$ log loss iff $K = 2$.
Metrics: Prec $= \frac{TP}{TP+FP}$, Rec(TPR) $= \frac{TP}{TP+FN}$, FPR $= \frac{FP}{FP+TN}$, FDR $= \frac{FP}{FP+TP} = 1 - \mathrm{Prec}$. F1 $= \frac{2PR}{P+R}$.
FP = type I (priority), FN = type II. Highest F1 $\neq$ always best (ignores other criteria). **ROC**: TPR vs FPR over diminishing τ. AUC = 1 perfect, 0.5 random

Kernels

Kernel trick: replace $\langle \boldsymbol{x}, \boldsymbol{x}' \rangle$ in $k \Rightarrow$ linear method to nonlinear. Reparam $\boldsymbol{w} = \Phi^\top \boldsymbol{\alpha} = \sum_i \alpha_i \phi(\boldsymbol{x}_i)$: a minimizer $\in \mathrm{span}\{\phi(\boldsymbol{x}_i)\}$ (representer; GD from $\boldsymbol{w}^0 = 0$ stays there) $\Rightarrow$ solve in $\boldsymbol{\alpha} \in \mathbb{R}^n$. Memory $O(np) \to O(n^2)$; poly k costs $O(n^2(d+m))$ vs naïve $O(n^2 d^m)$.
Valid kernel: $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle \Leftrightarrow$ (1) $k(\boldsymbol{x}, \boldsymbol{x}') = k(\boldsymbol{x}', \boldsymbol{x})$; (2) $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ PSD $\forall$ datasets. $k(\boldsymbol{x}, \boldsymbol{x}) < 0 \Rightarrow$ *invalid; but $k(\boldsymbol{x}, \boldsymbol{x}')$ may be < 0.* **Compositions**: valid $k_1 + k_2$; $k_1 k_2$; $c\,k_1$ & $k_1 + c$ ($c \geq 0$); $g(\boldsymbol{x}) k_1 g(\boldsymbol{x}')$ ($g: \mathcal{X} \to \mathbb{R}$); $f(k_1)$ if f non-neg Taylor coeffs (e.g. e^k, $\cosh k$); $k_1(\phi(\boldsymbol{x}), \phi(\boldsymbol{x}'))$. NOT valid $k_1 - k_2$; $\min$ & $\max(k_1, k_2)$;
Specific kernels: valid ($x, y \in \mathbb{R}/\mathbb{R}^d$) $xy = \langle x, y \rangle$; $g(\boldsymbol{x}) g(\boldsymbol{x}')$ (rank-1, $\phi = g$); $\cos(x - y)$ [$\phi = (\cos x, \sin x)$]; ($x, y \in \mathbb{R}_+$) $\min(x, y)$, $\frac{\min}{\max}$. NOT valid $\max(x, y)$ [$K = \left(\begin{smallmatrix} 2 & 2 \\ 2 & 1 \end{smallmatrix}\right) \not\succeq 0$].
Common Valid: **Linear** $\langle \boldsymbol{x}, \boldsymbol{x}' \rangle$: recovers linear regr./class. ($K = XX^\top$). **Polynomial** $(1 + \langle \boldsymbol{x}, \boldsymbol{x}' \rangle)^m$: $\phi =$ all monomials deg $\leq m$, dim $\binom{d+m}{m}$. Growth $O(m^d)$ fixed d, $O(d^m)$ fixed m.
Gaussian RBF $\exp(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2/\sigma^2)$ (∞-dim). Small σ: narrow $\Rightarrow$ *overfit.* large σ: nearly linear $\Rightarrow$ *underfit.* **Laplace** $\exp(-\|\boldsymbol{x} - \boldsymbol{x}'\|/\sigma)$.
Kernel Ridge Regression & PCA: $\min_{\boldsymbol{\alpha}} \|K\boldsymbol{\alpha} - \boldsymbol{y}\|^2 + \lambda \boldsymbol{\alpha}^\top K \boldsymbol{\alpha} \Rightarrow \boldsymbol{\alpha}^* = (K + \lambda I)^{-1} \boldsymbol{y}$, $f(\boldsymbol{x}) = \sum_i \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x})$.
Kernel PCA: kernel matrix $N \times N$ (not $d \times d$); $W_* = \Phi^\top A^\top$ (PCs in span of data); Gaussian $\Rightarrow \infty$-dim PCA.

Neural Networks

$f_{a\mathbf{W}}(\boldsymbol{x}) = a^K f_{\mathbf{W}}(\boldsymbol{x})$ (linear NN, K layers; NOT linear in $\mathbf{W}$; linear in $\boldsymbol{x}$).
Loss of deep NN is generally non-cvx. [$L(\mathbf{W})$ cvx $\Leftrightarrow$ only when optimizing linear last-layer with others fixed.]
Dense layer params: $m = n_{out}(n_{in} + 1)$
Activations: **1** ReLU $\max(0, z)$, $\phi' = \mathbf{1}_{z>0}$, cvx; **2** Sigmoid $\sigma = \frac{1}{1+e^{-z}}$, $\sigma' = \sigma(1 - \sigma)$; **3** Tanh $\frac{e^z - e^{-z}}{e^z + e^{-z}}$, $\tanh' = 1 - \tanh^2$; **4** Identity (only affine output). **Vanish. grad** (2,3); **non-diff. at 0** (1); **zero-centered** (3,4).
Backpropagation: Forward: $\boldsymbol{z}^l = W^l \boldsymbol{v}^{l-1}$; $\boldsymbol{v}^l = \phi(\boldsymbol{z}^l)$. Output: $\boldsymbol{\delta}^L = \nabla_{\boldsymbol{v}^L} \ell$; $\nabla_{W^L} \ell = \boldsymbol{\delta}^L (\boldsymbol{v}^{L-1})^\top$ Hidden: $\boldsymbol{\delta}^l = \phi'(\boldsymbol{z}^l) \odot \left[(W^{l+1})^\top \boldsymbol{\delta}^{l+1}\right]$; $\nabla_{W^l} \ell = \boldsymbol{\delta}^l (\boldsymbol{v}^{l-1})^\top$
Forward and backward both $O(L)$.
Initialization: **Zero init:** all activations $= 0 \Rightarrow$ ReLu: all gradients $= 0$. $\Rightarrow$ sigmoid: $\sigma(0) = 0.5$. **All-ones init:** all hidden units equal $\Rightarrow$ all rows of gradient equal. **He init:** $W_{ij} \sim \mathcal{N}(0, 2/p_k)$, variance $\approx$ constant across ReLU layers.
Too-small $\sigma \Rightarrow$ stalled training. Too-large $\sigma \Rightarrow$ exploding activations.
CNNs — Output Size: $l_{\text{out}} = \lfloor (l_{\text{in}} + 2p - f)/s \rfloor + 1$ (f=filter, p=padding, s=stride)
Parameters (F filters, $f \times f$, C_{in} input channels, with bias): $m = (f^2 \cdot C_{in} + 1) \times F$

(image size)*c ⇒ same # params
Channels/N ⇒ params/N.
Regularization: **Dropout:** zero out activations with prob $1-p$ at train. At test: **multiply weights by** p.
Weight decay = L2 reg: $\theta \leftarrow \theta(1-\alpha) - \eta\nabla\mathcal{L}_1 \Leftrightarrow$ L2 with $\lambda = \frac{\alpha}{2\eta}$.
Early stopping, batch norm.
Universal Approx.: 1 hidden layer of **fixed** width p: does **NOT** approximate all continuous functions. Need to let $p \to \infty$.

Clustering & PCA

K-Means: k centers $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k$. Cost $\hat{R}(\boldsymbol{\mu}) = \sum_{i=1}^n \min_{j\in\{1,\dots,k\}} \|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|_2^2$. Each $\|\cdot\|^2$ cvx, but $\arg\min_{\boldsymbol{\mu}} \hat{R}$ non-cvx, NP-hard (matrix factor. $X \approx WZ$, Z one-hot cols). Assign $z_i = \arg\min_j \|\boldsymbol{x}_i - \boldsymbol{\mu}_j\| \Rightarrow$ Voronoi partition of $\mathbb{R}^d$; model-based (extends to new points), unlike hierarchical/spectral.
Lloyd's: alternate (1) assign z_i; (2) $\boldsymbol{\mu}_j \leftarrow$ mean of cluster j. $O(nkd)$/iter. Cost never $\uparrow \Rightarrow$ converges, but to a **local** min only.
Init: uniform (bad if unbalanced); furthest-point (outlier-sensitive); **k-means++**: 1st center uniform, then $p(\boldsymbol{x}) \propto \min_j \|\boldsymbol{x} - \boldsymbol{\mu}_j\|^2$; expected cost within $O(\log k)$ of optimal.
Choosing k: train & val. cost both $\downarrow$ monotonically $\Rightarrow$ CV unusable; use elbow.
PCA: **Centered** data ($\sum_i \boldsymbol{x}_i = 0$). Linear, model-based dim. reduction $\boldsymbol{x}_i \in \mathbb{R}^d \to \boldsymbol{z}_i \in \mathbb{R}^k$, $k \ll d$ (manifold hypothesis). $W^*, z_1^*, \dots, z_n^* = \arg\min_{W^\top W = I, \boldsymbol{z}_i} \frac{1}{n}\sum_i \|\boldsymbol{x}_i - W\boldsymbol{z}_i\|_2^2$
Min. recon. error $\equiv$ max. proj. var. $\boldsymbol{w}^\top \Sigma \boldsymbol{w}$, $\Sigma = \frac{1}{n}\sum_i \boldsymbol{x}_i\boldsymbol{x}_i^\top = \frac{1}{n}X^\top X = \sum_{i=1}^d \lambda_i \boldsymbol{v}_i\boldsymbol{v}_i^\top$ (sym., PSD, spectral decomp.).
Solution: $W^* = [\boldsymbol{v}_1|\cdots|\boldsymbol{v}_k]$ top-k eigenvectors of Σ ($\lambda_1 \geq \dots \geq \lambda_d \geq 0$); $\boldsymbol{z}_i^* = W^{*\top}\boldsymbol{x}_i$; recon $= W^*W^{*\top}\boldsymbol{x}_i$. 1st PC = max-variance dir.; PCs orthonormal; sign of $\boldsymbol{v}_i$ ambiguous.
Variance along PC $j = \lambda_j$. Recon. error $C^{(k)} = \sum_{j=k+1}^d \lambda_j$; ($C^{(k-1)} - C^{(k)} = \lambda_k$, $C^{(d)} = 0$). Var. explained $= \sum_{j\leq k}\lambda_j / \sum_j \lambda_j$.
$P = WW^\top$ orthogonal projection: $P^2 = P$.
Choosing k: retain e.g. 95% var, or scree-plot elbow; CV if downstream supervised.
SVD: $X = USV^\top \Rightarrow n\Sigma = X^\top X = VS^\top SV^\top$; top-$k$ PCs = first k cols of V (right singular vectors); numerically stable. PCA = best rank-k approx of X (Eckart–Young).
Duplicate feature $\Rightarrow$ reduce dim by ≥ 1, zero recon. error.
Linear autoencoder (identity act., sq. recon. loss, bottleneck m) $\equiv$ PCA-m (Eckart–Young); depth irrelevant (linear act.).
PCA vs. k-means: both $\min \sum_i \|\boldsymbol{x}_i - W\boldsymbol{z}_i\|^2$; differ by constraint on $\boldsymbol{z}_i$ — PCA $W^\top W = I$, $\boldsymbol{z}_i \in \mathbb{R}^k$; k-means $\boldsymbol{z}_i \in \{\boldsymbol{e}_1, \dots, \boldsymbol{e}_k\}$.

Probabilistic Modeling

1. Model: **Discr.**: model only $p(y \mid x;\theta)$ — regr.: $Y \mid X{=}x \sim \mathcal{N}(f(x;\theta), \sigma^2)$; class.: $p(y \mid x;\theta) = \sigma(y\theta^\top x)$ (logistic; predict $Y{=}1$ iff $\theta^\top x > 0$). **Gen.**: model $p(x, y) = p(y)\,p(x \mid y)$ — NB: $\hat{y} = \arg\max_k p(k)\prod_i p(x_i \mid k)$ (cond. indep. given class).
GBC hierarchy (Σ_y:=Cov($x \mid Y{=}y$)): **QDA** per-class full Σ_y, unrestricted $\Rightarrow$ quadratic boundary; **LDA** $\Sigma_{+1}{=}\Sigma_{-1}$ shared $\Rightarrow$ linear; **Fisher's LDA** = LDA + $P(y){=}\frac{1}{2}$; **GNB** features of $\boldsymbol{x}$ i.i.d. given y $\Rightarrow \Sigma_y = \mathrm{diag}(\sigma_{y,1}^2, \dots)$.
2. Inference (fit θ): **MLE**: $\hat{\theta} = \arg\max_\theta \log p(\mathcal{D} \mid \theta)$. **MAP**: $\arg\max_\theta [\log p(\mathcal{D} \mid \theta) + \log p(\theta)]$. **Bayes**: full posterior $p(\theta \mid \mathcal{D}) \propto p(\mathcal{D} \mid \theta)p(\theta)$.
MAP = MLE iff uniform prior (Gaussian prior, $\sigma\to\infty$: MAP $\to$ MLE). MLE $\neq$ posterior mean in general. MLE minimizes $D_{\mathrm{KL}}(\hat{\pi}\|\nu)$ (empirical vs. model; $D_{\mathrm{KL}} \geq 0$, not sym.).
Neg. log-lik $\equiv$ loss: Gaussian noise $\Rightarrow$ squared loss; logistic model $\Rightarrow$ logistic regr. MAP Gaussian prior $\Rightarrow$ ridge; Laplace prior $\Rightarrow$ LASSO.
Known-MLEs: Bern: $\hat{\theta}{=}\bar{x}$; Gaus: $\hat{\mu}{=}\bar{x}$, $\hat{\Sigma}{=}\frac{1}{N}\sum(x_i{-}\hat{\mu})(x_i{-}\hat{\mu})^\top$; Exp: $\hat{\lambda}{=}\frac{N}{\sum x_i}$; Poisson: $\hat{\lambda}{=}\bar{x}$; Unif$(0,\theta)$: $\hat{\theta}{=}\max_i x_i$.
Conjugate prior: Beta-Bernoulli — $\theta \sim \mathrm{Beta}(\alpha,\beta)$, n_1 heads, n_0 tails $\Rightarrow \theta \mid \mathcal{D} \sim \mathrm{Beta}(\alpha{+}n_1, \beta{+}n_0)$; Beta$(1,1)$ = Unif$[0,1]$.
3. Decision (predict): **Bayes optimal predictor**: $f^*(x) = \arg\min_a \mathbb{E}[\ell(a, Y) \mid X{=}x]$ — best possible (no function class, min. gen. err.). Square loss $\Rightarrow$ cond. mean $\mathbb{E}[Y \mid X{=}x]$; 0-1 loss $\Rightarrow$ posterior mode $\arg\max_a p(a \mid x)$. With learned $\hat{\mathbb{P}}$: only an *estimate* of the optimal rule.

Expectation Maximization & GMMs

GMM: $p(\boldsymbol{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \Sigma_k)$; $\sum_k \pi_k = 1$. Latent: $z_i \sim \mathrm{Cat}(\boldsymbol{\pi})$; $\boldsymbol{x}_i \mid z_i = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$.
Covariance structures: **Spherical** $\Sigma{=}\sigma^2 I$ (circles; 1 param); **Diagonal** (axis-aligned; d params); **Full** (any orient.; $d(d{+}1)/2$ params).
Complete-data log-lik: $\log p(X, Z \mid \theta) = \sum_i \sum_k \mathbf{1}_{z_i=k}[\log \pi_k + \log \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \Sigma_k)]$
EM Algorithm: **E-step** — responsibilities (soft assignments): $\gamma_k(\boldsymbol{x}_i) = \frac{\pi_k\, p(\boldsymbol{x}_i \mid z_i{=}k)}{\sum_j \pi_j\, p(\boldsymbol{x}_i \mid z_i{=}j)}$ **Q-function:** $Q(\theta;\theta^t) = \mathbb{E}_Z[\log p(X, Z \mid \theta) \mid X, \theta^t] = \sum_i \sum_k \gamma_k(\boldsymbol{x}_i) \log p(\boldsymbol{x}_i, z_i{=}k;\theta)$
M-step — maximize Q in closed form (**no step size**, unlike GD): $\pi_k^{(t+1)} = \frac{1}{N}\sum_i \gamma_k(\boldsymbol{x}_i)$
Convergence: soft vs. hard: **Soft:** $\gamma_k \in (0,1)$ is the **true** posterior $\Rightarrow$ bound **tight** $\Rightarrow$ log-lik **never** $\downarrow$ (monotone). **Hard:** γ_k rounded to $\{0,1\}$ ($\arg\max_k$) $\Rightarrow$ bound **loose** $\Rightarrow$ **no** monotonicity; cheaper (no sum over z); $\equiv k$-means if $\Sigma_k{=}\sigma^2 I$. **Both:** log-lik **non-cvx** for $K{>}1$ (label-perm. symmetry, overlapping comps.) $\Rightarrow$ **local** max only, init-dependent.
Semi-supervised: fix z_i for labeled pts., unlabeled latent $\Rightarrow$ EM still applies.
Degeneracies & Choosing K: If $\pi_k^{(0)}{=}0$: $\gamma_k(\boldsymbol{x}_i) \propto \pi_k{=}0\ \forall i \Rightarrow \pi_k^{(t)}{=}0$ for all t (stays 0 forever). If $\boldsymbol{\mu}_0^{(0)} = \boldsymbol{\mu}_1^{(0)}$: assignments stay at prior $\Rightarrow$ params stay equal (degenerate init; cf. k-means bad init).
Train. log-lik $\uparrow$ monotonically in K (more components = more capacity) $\Rightarrow$ **cannot** choose K by train MLE; use CV.

Language Models

Goal: model $P(\boldsymbol{x}) = P(x_1, \dots, x_m)$ over token sequences. **Autoregressive factorization** (chain rule): $P(x_1, \dots, x_m) = \prod_{t=1}^m P(x_t \mid x_1, \dots, x_{t-1})$
Context window k: $P(x_t | x_{t-k:t-1};\theta) = \mathrm{Cat}(\mathrm{softmax}(f(x_{t-k:t-1};\theta)))$
Training loss (NLL, self-supervised — no labels needed): $\min_\theta -\sum_t \log P(X_t = x_t | X_{t-k:t-1};\theta)$
Pipeline: **pre-train** on massive corpus (self-sup.) $\to$ **fine-tune** on small task-specific set. Archit.: **decoder-only** (causal), enc-dec, or enc-only — *not* exclusively enc-dec.
Input: **tokens** (words/sub-words); **one-hot** sparse, no semantics; **embeddings** $W_e \in \mathbb{R}^{N\times d}$ learnable *lookup* (not via MLP, unlike q,k,v), dense + semantic. **Positional emb.** $z_i = w_{e,i} + w_{p,i}$: self-attn. is **permutation-invariant** (input = set) $\Rightarrow$ required for order, else bag-of-tokens.
RNN vs. attn.: RNN $h_t = g(h_{t-1}, x_t)$: long-range info lost, sequential $\Rightarrow$ no parallelism. Attn.: any two positions linked **directly** regardless of distance, **parallel** over seq.; but *no* param. reduction, and $O(k^2)$ cost in context k.
Self-Attention: For each token i: $q_i = z_iW^Q$, $k_i = z_iW^K$, $v_i = z_iW^V$ ($W^Q, W^K, W^V \in \mathbb{R}^{d\times d_k}$). $z_i' = \sum_j s_{ij} v_j$; $s_{ij} \propto \exp\left(\frac{q_i k_j^\top}{\sqrt{d_k}} + m_{ij}\right)$
Causal mask: $m_{i,j} = -\infty$ if $j{>}i$, else 0 (can't peek at future). **Matrix:** $Z' = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + M\right)V$. $\sqrt{d_k}$: prevents large dot products saturating softmax.
Multi-head: h heads, own W_r^Q, W_r^K, W_r^V, concat: MultiHead(Z) = Concat$(Z_1', \dots, Z_h')\,W^O$; each head a different relation type.
Block ($\times n$): masked multi-head self-attn. $\to$ resid. + LayerNorm (prevents vanish. grad.) $\to$ per-token FFN (no cross-token mixing) $\to$ resid. + LayerNorm.
Decoding: train: all tokens in 1 forward pass (causal mask). Inference: m tokens $\Rightarrow$ m sequential passes (can't parallelize).

Math Toolbox

Gradients: $\nabla_{\boldsymbol{w}}(\boldsymbol{a}^\top\boldsymbol{w}) = \boldsymbol{a}$; $\nabla_{\boldsymbol{w}}(\boldsymbol{a}^\top\boldsymbol{w})^2 = 2(\boldsymbol{a}^\top\boldsymbol{w})\boldsymbol{a}$; $\nabla_{\boldsymbol{w}}(\boldsymbol{c}^\top A\boldsymbol{w}) = A^\top\boldsymbol{c}$; $\nabla_{\boldsymbol{w}}(\boldsymbol{w}^\top A\boldsymbol{w}) = (A{+}A^\top)\boldsymbol{w}$ (= $2A\boldsymbol{w}$ **only if** A sym.); $\nabla_{\boldsymbol{w}}\|A\boldsymbol{w}{-}\boldsymbol{b}\|_2^2 = 2A^\top(A\boldsymbol{w}{-}\boldsymbol{b})$
Norms: $\|\boldsymbol{a} \pm \boldsymbol{b}\|_2^2 = \|\boldsymbol{a}\|_2^2 \pm 2\langle \boldsymbol{a}, \boldsymbol{b}\rangle + \|\boldsymbol{b}\|_2^2$; $\nabla_{\boldsymbol{w}}\|\boldsymbol{w}\|_2^2 = 2\boldsymbol{w}$; $\nabla_{\boldsymbol{w}}\|\boldsymbol{w}\|_2 = \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$ ($\boldsymbol{w}{\neq}\boldsymbol{0}$); $\nabla_{\boldsymbol{w}}\|\boldsymbol{w}\|_1 = \mathrm{sign}(\boldsymbol{w})$ (element-wise; subgrad. $[-1,1]$ at 0)
Chain rule: $\nabla_{\boldsymbol{w}} f(g(\boldsymbol{w})) = (\nabla_{\boldsymbol{w}} g)^\top \nabla_g f$; scalar g: $f'(g(\boldsymbol{w}))\,\nabla_{\boldsymbol{w}} g(\boldsymbol{w})$.