Background

Linear Algebra
$\|\mathbf{x}\|_p = (\sum |x_i|^p)^{1/p}$ $\quad \|\mathbf{x}\|_\infty = \max|x_i|$
$\mathrm{tr}(\mathbf{Axx}^T) = \mathbf{x}^T\mathbf{Ax}$
$|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$ $\quad |\mathbf{A}^m| = |\mathbf{A}|^m$
$(\mathbf{A}+\mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1}+\mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1}$
$(\mathbf{A}+\mathbf{B})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}^{-1}$
$\mathbf{U}(\mathbf{VU}+\mathbf{I})^{-1} = (\mathbf{UV}+\mathbf{I})^{-1}\mathbf{U}$
$\mathbf{I}-\mathbf{A}(\mathbf{I}+\mathbf{A})^{-1} = (\mathbf{I}+\mathbf{A})^{-1}$

Probability
$B(a,b) = \Gamma(a)\Gamma(b)\Gamma^{-1}(a+b)$
$\Gamma(a) = \int_0^\infty e^{-x}x^{a-1}\,dx$
$\mathrm{Ber}(x|\theta) = \theta^x(1-\theta)^{1-x}$ $\quad 0\le\theta\le1$
$p_Y(y) = p_X(g^{-1}(y))\left|\det\frac{\partial g^{-1}(y)}{\partial y}\right|$
$\mathbb{E}_{Y|X}[Y] = \mathbb{E}_Y[Y|X]$ | $\mathbb{E}_Y[\mathbb{E}_X[X|Y]] = \mathbb{E}_X[X]$
$\mathbb{E}_{X,Y}[f(X,Y)] = \mathbb{E}_X\mathbb{E}_{Y|X}[f(X,Y)|X]$
$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X\mid Y)] + \mathrm{Var}(\mathbb{E}[X\mid Y])$
$\mathbb{V}[X+Y] = \mathrm{Var}[X]+\mathrm{Var}[Y]+2\mathrm{Cov}(X,Y)$
$\mathrm{Cov}(\mathbf{X},\mathbf{Y}) = \mathbb{E}[\mathbf{XY}^T] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{Y}]^T$
$\mathrm{Cov}(A\mathbf{X}+c, B\mathbf{Y}+d) = A\mathrm{Cov}(\mathbf{X},\mathbf{Y})B^T$
$\mathcal{N}(x|\mu,\Sigma) = \frac{exp(-\frac12(x-\mu)^T\Sigma^{-1}(x-\mu))}{(2\pi)^{D/2}|\Sigma|^{1/2}}$
$X = \Sigma^{1/2}\mathcal{N}(0,1)+\mu \sim \mathcal{N}(\mu,\Sigma)$
$Y = MX+b \sim \mathcal{N}(M\mu+b, M\Sigma M^T)$
$X,Y \overset{\text{iid}}{\sim} \mathcal{N}:\quad X+Y\sim\mathcal{N}(\mu+\mu',\Sigma+\Sigma')$

Conditional Gaussian
$P(\begin{bmatrix}\mathbf{X}\\\mathbf{Y}\end{bmatrix}) = \mathcal{N}(\begin{bmatrix}\mathbf{X}\\\mathbf{Y}\end{bmatrix}; \begin{bmatrix}\mu_1\\\mu_2\end{bmatrix}, \begin{bmatrix}\Sigma_{11}&\Sigma_{12}\\\Sigma_{21}&\Sigma_{22}\end{bmatrix})$
$p(\mathbf{Y}|\mathbf{X}=\mathbf{x}) = \mathcal{N}(\mu,\Sigma)$
$\cdot\ \mu = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{x}-\mu_{\mathbf{X}})$
$\cdot\ \Sigma = \Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$

Inequalities and Estimators
Jensen: $log(\sum_i\lambda_i^{(\ge0)}x_i) \ge \sum_i\lambda_i log(x_i)$
Chebyshev: $\mathbb{P}(|\hat X - X|\ge\varepsilon) \le \frac{MSE[\hat X]}{\varepsilon^2}$
Estimators: *Unbiased:* $\mathbb{E}[\hat\theta] = \theta^\star$
Consistent: $\mathbb{P}(|\hat\theta-\theta^\star|<\varepsilon)\to 0$ convP
Asymp Normal: $(\hat\theta-\theta^\star)\hat{se}^{-1}\sim\mathcal{N}(0,1)$
Rao-Cra.: $\mathbb{E}_{x|\theta}[(\theta-\hat\theta)^2] \ge \frac{(\frac{\partial}{\partial\theta}b_{\hat\theta}+1)^2}{\mathbb{E}_{x|\theta}[\Lambda^2]} + b_{\hat\theta}^2$
$b_{\hat\theta} = \mathbb{E}_{x|\theta}[\hat\theta]-\theta$ $\quad \Lambda = \frac{\partial}{\partial\theta}\log p(x|\theta)$
$\mathbb{E}_{x|\theta}[\Lambda] = 0 \to \mathbb{E}_{x|\theta}[\Lambda\hat\theta] = \frac{\partial}{\partial\theta}b_{\hat\theta}+1$
$\to \mathrm{Cov}(\Lambda,\hat\theta) \to$ Cauchy
$\mathrm{Var}[\hat\theta] \ge \mathcal{I}_n(\theta)^{-1} = -\mathbb{E}[\frac{\partial^2\log p[\mathcal{X}_n|\theta]}{\partial\theta^2}]^{-1}$
Efficiency of $\hat\theta$: $e(\theta_n) = \frac{1}{\mathrm{Var}[\hat\theta_n]\mathcal{I}_n(\theta)}$
$\hat\theta_{JS} = \left(1-\frac{(d-2)\sigma^2}{\|y\|^2}\right)y$

Derivatives
$\frac{\partial}{\partial\mathbf{x}}(\mathbf{b}^\top\mathbf{x}) = \frac{\partial}{\partial\mathbf{x}}(\mathbf{x}^\top\mathbf{b}) = \mathbf{b}$
$\frac{\partial}{\partial\mathbf{x}}(\mathbf{x}^\top\mathbf{Ax}) = (\mathbf{A}^\top+\mathbf{A})\mathbf{x}$
$\frac{\partial}{\partial\mathbf{X}}(\mathbf{c}^\top\mathbf{Xb}) = \mathbf{cb}^\top$ $\quad \frac{\partial}{\partial\mathbf{X}}(\|\mathbf{X}\|_F^2) = 2\mathbf{X}$
$\frac{\partial}{\partial\mathbf{x}}\|\mathbf{x}\|_2 = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$ | $\frac{\partial}{\partial\mathbf{x}}\|f(\mathbf{x})\|_1 = \frac{\partial f(\mathbf{x})}{\partial\mathbf{x}}^T\mathrm{sgn}(\mathbf{x})$
$\frac{\partial}{\partial\mathbf{x}}(\|\mathbf{Ax}-\mathbf{b}\|_2^2) = 2\mathbf{A}^\top(\mathbf{Ax}-\mathbf{b})$
$\frac{\partial}{\partial\mathbf{X}}(|\mathbf{X}|) = |\mathbf{X}|\cdot\mathbf{X}^{-1}$, $\quad |\mathbf{X}|^{-1} = |\mathbf{X}^{-1}|$
$\frac{\partial}{\partial\mathbf{X}}f(\mathbf{X})^\top = \frac{\partial f(\mathbf{X})}{\partial\mathbf{X}}^T$ | $\frac{\partial}{\partial\mathbf{X}}\mathrm{tr}\,f(\mathbf{X}) = \mathrm{tr}\,\frac{\partial f(\mathbf{X})}{\partial\mathbf{X}}$
$\frac{\partial}{\partial\mathbf{X}}\det f(\mathbf{X}) = \det f(\mathbf{X})\,\mathrm{tr}(f(\mathbf{X})^{-1}\frac{\partial f(\mathbf{X})}{\partial\mathbf{X}})$
$\frac{\partial}{\partial\mathbf{X}}f(\mathbf{X})^{-1} = -f(\mathbf{X})^{-1}\frac{\partial f(\mathbf{X})}{\partial\mathbf{X}}f(\mathbf{X})^{-1}$

Quadratic Forms
$\mathbf{x}^TA\mathbf{x}+2\mathbf{b}^T\mathbf{x}+c = (\mathbf{x}+A^{-1}\mathbf{b})^TA(\mathbf{x}+A^{-1}\mathbf{b}) - \mathbf{b}^TA^{-1}\mathbf{b}+c$,
$ax^2+bx+c = (x+\frac{b}{2a})^2 - (\frac{b}{2a})^2+c$

Information Theory
$\mathrm{H}[p] = \mathbb{E}_{\mathbf{x}\sim p}[-\log p(\mathbf{x})]$
$\mathrm{H}[p\|q] = \mathbb{E}_{\mathbf{x}\sim p}[-\log q(\mathbf{x})]$
$\mathrm{KL}[p\|q] = \mathrm{H}[p\|q]-\mathrm{H}[p]$
$\mathrm{KL}[p\|q] = \mathbb{E}_{\theta\sim p}\left[\log\left(\frac{p(\theta)}{q(\theta)}\right)\right]$
$\mathrm{KL}[p\|q]\ne\mathrm{KL}[q\|p]\ge0$
$\mathrm{H}[\mathbf{X}] = \mathbb{E}_{\mathbf{X}\sim p}[-\log p(\mathbf{X})]$
$\mathrm{H}[\mathbf{X}|\mathbf{Y}=y] = \mathbb{E}_{\mathbf{X}\sim p(\cdot|y)}[-\log p(\mathbf{X}|y)]$
$\mathrm{H}[\mathbf{X}|\mathbf{Y}] = \mathbb{E}_y[\mathrm{H}[\mathbf{X}|\mathbf{Y}=y]]$
$\mathrm{H}[\mathbf{X}|\mathbf{Y}] = \mathrm{H}[\mathbf{Y}|\mathbf{X}]+\mathrm{H}[\mathbf{X}]-\mathrm{H}[\mathbf{Y}]$
$\mathrm{H}[\mathbf{X},\mathbf{Y}] = \mathbb{E}_{(\mathbf{X},\mathbf{Y})\sim p(\cdot,\cdot)}[-\log p(\mathbf{X},\mathbf{Y})]$
$\mathrm{I}[\mathbf{X};\mathbf{Y}] = \mathrm{H}[\mathbf{X}]-\mathrm{H}[\mathbf{X}|\mathbf{Y}]\ge0$
$\mathrm{I}[\mathbf{X};\mathbf{Y}|\mathbf{Z}] = \mathrm{I}[\mathbf{X};\mathbf{Y},\mathbf{Z}]-\mathrm{I}[\mathbf{X};\mathbf{Z}]$
$\mathrm{H}(\mathcal{N}(\mu,\Sigma)) = \frac12\ln(\det(2\pi e\Sigma))$
$\mathrm{KL}(\mathcal{N}(a,A)\|\mathcal{N}(b,B)) = \frac12(\mathrm{tr}(B^{-1}A) + (a-b)^TB^{-1}(a-b) - d + \ln(\frac{\det B}{\det A}))$

Risks
Expected Risk: $R(f) = P(f(X)\ne y)$
$\mathcal{R}(f) = \sum_{y\le k}P(y)\mathbb{E}_{P(x|y)}[1_{f(x)\ne y}|Y=y]$
Empirical Risk Minimizer (ERM) $\hat f$:
$\hat f\in\arg\min_{f\in\mathcal{H}}\hat R(\hat f,\mathcal{D}^{train})$
$\hat R(\hat f,\mathcal{D}^{train}) = \frac1n\sum_{i=1}^n\mathcal{L}(Y_i,\hat f(X_i))$
$\hat R(\hat f,\mathcal{D}^{test}) = \frac1m\sum_{i=n+1}^{n+m}\mathcal{L}(Y_i,\hat f(X_i))$

Loss Fcts: $\mathcal{L}(y,z)$ $\quad z=w^\top x$
$\mathcal{L}^{0/i} = \mathbb{I}[\mathrm{sign}(z)\ne y]$
$\mathcal{L}^{\text{hinge}} = \max(0,1-yz)$ for SVM's
$\mathcal{L}^{\text{percep}} = \max(0,-yz)$
$\mathcal{L}^{\text{logistic}} = \log(1+\exp(-yz))$
$\mathcal{L}^{\exp} = \exp(-yz)$ for AdaBoost
$\mathcal{L}^{\text{CE}} = -[y'\log z' + (1-y')\log(1-z')]$
$y' = \frac{1+y}{2}$, $\quad z' = \frac{1+z}{2}$

Optimization
GDM: $\theta^{(t+1)}\leftarrow\theta^{(t)}-\eta\nabla_\theta\mathcal{L}+\mu(\theta^{(t)}-\theta^{(t-1)})$
GD: $\theta^{(t+1)}\leftarrow\theta^{(t)}-\eta\nabla_\theta\mathcal{L}$
SGD: $\theta^{(t+1)}\leftarrow\theta^{(t)}-\eta\nabla\mathcal{L}(\theta^{(t)},x_i,y_i)$
NGD: $\theta^{(t+1)}\leftarrow\theta^{(t)}-\eta(\nabla_\theta^2\mathcal{L})^{-1}\nabla_\theta\mathcal{L}$
$\to f(x+t)\approx f(x)+tf'(x)+\frac12f''(x)t^2=0$

Parametric Density Estimation
Assume prior $\mathbb{P}(\theta)$,
Likelihood: $\mathbb{P}[\mathcal{X}|\theta] = \prod_{i\le n}p(x_i|\theta)$
$\hat\theta_{MLE} = \arg\max_\theta\mathbb{P}[\mathcal{X}|\theta]$
$\hat\theta_{MAP} = \arg\max_\theta[P(\theta|\mathcal{X}) = P(\mathcal{X}|\theta)P(\theta)]$
Solve $\nabla_\theta logP(\mathcal{X}|\theta)P(\theta) = 0$

1-D Gaussian Bayesian learning
$X|\theta\sim\mathcal{N}(\theta,\sigma^2)$ $\quad \theta\sim\mathcal{N}(m_0,s_0^2)$
$\theta|X\sim\mathcal{N}(\mu_n,\sigma_n^2)$
$\sigma_n^2 = \frac{\sigma^2s_0^2}{ns_0^2+\sigma^2}$, $\quad \mu_n = \frac{ns_0^2\bar x+m_0\sigma^2}{ns_0^2+\sigma^2}$

Recursive Bayesian density learning
$\mathcal{X}^n = x_{1:n}: p(\theta|\mathcal{X}^n) = \frac{p(x_n|\theta)p(\theta|\mathcal{X}^{n-1})}{\int p(x_n|\theta)p(\theta|\mathcal{X}^{n-1})d\theta}$

Frequentist vs Bayesian
Bayes: priors, distributions, needs efficient integration, adds regularization term. Frequentist: no priors, point estimate, requires only differentiation methods. MLE are consistent, equivariant, asymptotically normal, asymptotically efficient (no efficient for finite samples).

Data Types
monadic: $X:O\to\mathbb{R}^d$ dyadic: $X:O_1\times O_2\to\mathbb{R}^d$. pairwise: $X:O_1\times O_1\to\mathbb{R}^d$ polyadic data: $X:O_1\times O_2\times O_3\to\mathbb{R}^d$ nominal = qualitative (sweet, sour ...), ordinal = absolute order, quantitative = numbers

Regression
Model of data: $\mathbf{Y} = \mathbf{X}\beta^\star+\varepsilon$
$\mathbf{X}\in\mathbb{R}^{(d+1)\times n}$ $\quad \beta\in\mathbb{R}^{d+1}$ $\quad \varepsilon\sim\mathcal{N}(\mathbf{0},\mathbb{I}\sigma^2)$
$\mathbf{Y}|\mathbf{X},\beta,\sigma^2\sim\mathcal{N}(\mathbf{Y};\mathbf{X}^T\beta,\mathbb{I}_{(d+1)}\sigma^2)$

MLE: Ordinary Least Squares
OLSE is unbiased, orthogonal projection with lowest variance. differentiate wrt β.
$\mathcal{L} = \mathrm{RSS}(\beta) = \sum_{i=1}^n(y_i-x_i^T\beta)^2 = (\mathbf{y}-\mathbf{X}\beta)^2$
Estimator: $\hat\beta^{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
Prediction: $\hat y = \mathbf{X}\hat\beta = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

MAP: Ridge Regression (L^2 penalty)
Penalize energy in β. *Prior:* $\beta\sim\mathcal{N}(0,\lambda^{-1}\mathbb{I})$
Loss: $\mathcal{L} = (\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)+\lambda\beta^T\beta$
Estimator: $\hat\beta^{\text{ridge}} = (\mathbf{X}^T\mathbf{X}+\lambda\mathbb{I})^{-1}\mathbf{X}^T\mathbf{y}$

MAP: Lasso (L^1 penalty)
Penalize full β. Lasso has no closed form.
$\beta\sim Lapl(0,\lambda^{-1}) = \frac\lambda2exp(-\lambda|\beta|)$
$\mathcal{L} = \sum_{i=1}^n(y_i-x_i^T\beta)^2+\lambda\sum_{j=1}^d|\beta_j|$
$= (\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)+\lambda||\beta||_1$
Bayesian view: $Y|(X,\beta)\sim\mathcal{N}(x^T\beta,\sigma^2I)$

d-Dim Bayesian Linear Regression
Prior: $\beta\sim\mathcal{N}(\mu_0,\Lambda^{-1})$
Likelihood: $Y|\beta,X,\sigma\sim\mathcal{N}(X\beta,\sigma_n^2\mathbb{I})$
Posterior: $\beta|\mathbf{X},\mathbf{y}\sim\mathcal{N}(\mu,\Sigma)$
$\cdot\ \Sigma = (\sigma_n^{-2}\mathbf{X}^T\mathbf{X}+\Lambda)^{-1}$
$\cdot\ \mu = \Sigma(\Lambda\mu_0+\sigma_n^{-2}\mathbf{X}^T\mathbf{y})$

Nonlinear Regression
Idea: Add feature space transformation, kernel to compute inner product. Suppose:
$\beta\sim\mathcal{N}(\mathbf{0},\Lambda^{-1})$ $\quad \varepsilon\sim\mathcal{N}(\mathbf{0},\sigma_n^2\mathbb{I}_d)$
$\mathbf{Y} = \mathbf{X}\beta+\varepsilon\sim\mathcal{N}(\mathbf{0},\mathbf{X}\Lambda^{-1}\mathbf{X}^T+\sigma_n^2\mathbb{I}_d)$

Kernels
Kernel: $k(x_i,x_j) = \phi(x_i)\Lambda^{-1}\phi(x_j)^T$
Similarity based reasoning.
Gram Matrix: $K = k(\mathbf{x}_i,\mathbf{x}_j)$, $\quad 1\le i,j\le n$
$\cdot\ k(\mathbf{x},\mathbf{x}') = k(\mathbf{x}',\mathbf{x})$ $\cdot\ k(\mathbf{x},\mathbf{x}')$ pos.semi-def.
If k_1,k_2 kernels, $c\in\mathbb{R}_{>0}$, $\mathbf{A}^{psd}$, $p_{\text{pos-coeff}}$:
$k(\mathbf{x},\mathbf{x}') = k_1(\mathbf{x},\mathbf{x}')+k_2(\mathbf{x},\mathbf{x}') = \mathbf{x}^T\mathbf{A}\mathbf{x}'$
$= k_1(\mathbf{x},\mathbf{x}')\cdot k_2(\mathbf{x},\mathbf{x}') = c\cdot k_1(\mathbf{x},\mathbf{x}')$
$= p(k_1(\mathbf{x},\mathbf{x}')) = f(\mathbf{x})k_1(\mathbf{x},\mathbf{x}')f(\mathbf{x}')$

$k(\mathbf{x},\mathbf{x}') = \phi(x)^T\phi(x') = (1+\mathbf{x}^T\mathbf{x}')^m$
$= \tanh(\alpha\mathbf{x}^T\mathbf{x}'+c)$
$= \sigma^2\exp(-\frac{2\sin(p^{-1}\pi||\mathbf{x}-\mathbf{x}'||_2^2)}{l^2})$
$= \exp(-||\mathbf{x}-\mathbf{x}'||_1\,l^{-1})$
$= \exp(-||\mathbf{x}-\mathbf{x}'||_2^2\,(2l^2)^{-1})$
RBF: $\phi_j(x) = \exp(-\frac{||x||_2^2}{2})\prod_{i=0}^dx^{j_i}(j_i!)^{-\frac12}$
$\uparrow$ Lengthscale, smoother fcts.

Gaussian Process Regression
Applying a kernel, we get:
$\mathbf{Y} = \Phi\beta+\varepsilon\sim\mathcal{N}(\mathbf{0},\Phi\Lambda^{-1}\Phi^T+\sigma_n^2\mathbb{I}_d) = \mathcal{N}(\begin{bmatrix}\mathbf{y}\\y_*\end{bmatrix}|\mathbf{0},\begin{bmatrix}\mathbf{K}+\sigma^2\mathbb{I} & \mathbf{k}\\\mathbf{k}^\mathbf{T} & k(x_*,x_*)+\sigma^2\end{bmatrix})$

Gaussian Process Prediction
Given $\mathcal{GP}(\mu,K)$,
$p(y_*|x_*,\mathbf{X},\mathbf{y}) = \mathcal{N}(\tilde\mu,\tilde\sigma^2)$,
$\cdot\ \tilde\mu = \mu(x_*)+\mathbf{k}^T(\mathbf{K}+\sigma_n^2\mathbb{I})^{-1}(\mathbf{y}-\mu(\mathbf{X}))$,
$\cdot\ \tilde\sigma^2 = k(x_*,x_*)-\mathbf{k}^T(\mathbf{K}+\sigma^2\mathbb{I})^{-1}\mathbf{k}$
$\cdot\ \mathbf{k} = k(x_*,\mathbf{X})$ $\quad \mathbf{K}_{ij} = k(x_i,x_j)$
$\cdot\ \tilde\sigma_{ij}^2 = k(x_i,x_j)-\mathbf{k}_i^T(\mathbf{K}+\sigma^2\mathbb{I})^{-1}\mathbf{k}_j$

Causality
Counterfactual Invariance: A function f is invariant if $f(X(w)) = f(X(w'))\,\forall w,w'$, $\downarrow$ bias from spurious correlations.
Confounding: A hidden variable influences W and X, $\Rightarrow$ spurious correlation with Y. **Selection Bias:** A hidden variable S filters the training data based on W and X, inducing non-causal associations.
If f is counterf. invar.:
anti-causal scenario: $f(X)\perp W|Y$.
causal scenario (no selection): $f(X)\perp W$.
causal scenario (no confounding): $Y\perp X\mid W, X_{\perp W}$ and $f(X)\perp W\mid Y$.
A set of variables Z **d-separates** X and Y in a DAG $\mathcal{G}$ if all paths between X and Y are blocked by Z: $X\perp Y|Z$. A path is blocked if: **Collider:** $A\to B\leftarrow C$ and neither B nor its descendants are in Z. **Chain:** $A\to B\to C$. **Fork:** $A\leftarrow B\to C$ where $B\in Z$.

Algos
K-Means $J = \sum_{x\in\mathcal{X}}||x-\mu_{c(x)}||^2$
PCA proj. maximum variance subspace. top d eigenv. of $S = \frac1n\sum_{i=1}^n(x_i-\overline X)(x_i-\overline X)^T$
EM fit GMMs $(\sum_{k=1}^K\pi_k\mathcal{N}(x|\mu_k,\Sigma_k))$ by max. likelihood. Reaches local optimum.
Latent variable: $M_{xc} = 1\{c \text{ generated } x\}$
$P(\mathcal{X},M|\theta) = \prod_x\prod_{c=1}^k(\pi_cP(x|\theta_c))^{M_{xc}}$
$\gamma_{xc} = \mathbb{E}[M_{xc}|\mathcal{X},\theta^{(j)}] = \frac{\pi_c\mathcal{N}(\mathbf{x};\mu_c,\Sigma_c)}{\sum_{j=1}^k\pi_j\mathcal{N}(\mathbf{x};\mu_j,\Sigma_j)}$
$\mu_c^{(j+1)} = \frac{\sum_{c\in\mathcal{X}}\gamma_{xc}x}{\sum_{c\in\mathcal{X}}\gamma_{xc}}$ $\quad \pi_c^{(j+1)} = \frac{1}{|\mathcal{X}|}\sum_{c\in\mathcal{X}}\gamma_{xc}$
$(\sigma_c^2)^{(j+1)} = \frac{\sum_{c\in\mathcal{X}}\gamma_{xc}(x-\mu_c)^2}{\sum_{c\in\mathcal{X}}\gamma_{xc}}$

Perceptron Bound: $\frac{\max_{i\in\tilde{\mathcal{X}}^{mc}}\|\tilde x_i\|^2\|\hat a\|}{(\min_{i\in\tilde{\mathcal{X}}^{mc}}(\hat a^\top\tilde x_i))^2}$

Bias-Variance tradeoff
Bias$(\hat f) = \mathbb{E}[\hat f]-f$
Var$(\hat f) = \mathbb{E}[(\hat f-\mathbb{E}[\hat f])^2]$

Squared Error Decomposition
$\mathbb{E}_D\mathbb{E}_{X,Y}[(\hat f(X)-Y)^2] =$
$\mathbb{E}_{X,Y}[(\mathbb{E}_{Y|X}[Y]-Y)^2]$ (noise var)
$+\mathbb{E}_X\mathbb{E}_D[(\hat f_D(X)-\mathbb{E}_D[\hat f(X)])^2]$ (var.)
$+\mathbb{E}_X[(\mathbb{E}_D[\hat f_D(X)]-\mathbb{E}_{Y|X}[Y])^2]$ (bias^2)
With $\mathbb{E}_{Y|X}[Y]$ the expected label and $\mathbb{E}_D[\hat f(X)]$ the expected classifier.

p-value
p-value $= \inf\{\alpha: T(X^n)\in\{x|T(x)\ge c\}\}$ likelihood to accept H_0. it is the least probable threshold for rejecting the H_0.

Statistical Learning and Validation
Find $f: X\to Y$ to minimize expected risk by approximation with empirical risk.

K-Fold Cross Validation

Partition data Z into K equa. subsets: $\mathcal{Z} = \mathcal{Z}_1 \bigcup \mathcal{Z}_2 \bigcup \cdots \bigcup \mathcal{Z}_K$, $\mathcal{Z}_\mu \bigcap \mathcal{Z}_\nu = \emptyset$
$|\mathcal{Z}_k| \approx n\frac{K-1}{K}$ # of training samples. Learn: $\hat f^{-\nu}(x) = \arg\min_{f\in\mathcal F} \frac{\sum_{i \notin \mathcal Z_\nu} \mathcal L(y_i, f(x_i))}{|\mathcal Z - \mathcal Z_\nu|}$
$\hat R^{cv}(\mathcal A) = \frac1n \sum_{i\le n} \mathcal L(y_i, \hat f^{-\kappa(i)}(x_i))$
Underfits because smaller dataset.
Leave-one-out: $K=n$ (unbiased but var can be large from correlated datasets)
Bootstrapping $\mathcal Z^* = \{\mathcal Z_1^*, \cdots \mathcal Z_B^*\}$, of same size as original, drawn with replacement. a sample to have appears in bootstrap with prob: $1-(1-n^{-1})\approx 0.632$. So if we compute the ERM on $\mathcal Z$ we could get 63% accuracy by memorization. Over-confident (shows too small bias)!
Leave-one-out/out of bucket error: compensates by computing the ERM where no memorization was for specific sample. E.g., for classification, like cross-validation:
$\hat{\mathcal R}(\mathcal A) = \frac1B\sum_{b=1}^B \sum_{z_i\notin \mathcal Z^{*b}} \frac{\mathbb I_{c(x_i)\ne y_i}}{B - |\mathcal Z^{*b}|}$ $\hat R_{0.632} = 0.368\hat R(A(Z)) + 0.632\hat R_{bs}$
Wald Test: $W = \frac{\hat\theta-\theta_0}{\text{s.e.}(\hat\theta)}$

Bayesian Neural Networks (BNN)
NN: no uncertainty quantification, overconfident, adversarial examples, poor generalization for domain shifts. BNN: Using $p(w)$ and $p(D|w)$, approx. poster. by variational infer. (min rev KL).
$\sigma \leftarrow \sigma - \alpha_t \left(\epsilon^\top \frac{\partial}{\partial w}F(w,\theta) + \frac{\partial}{\partial\sigma}F(w,\theta)\right)$

Information-based Transductive Lear.

ITL selects x_n that maximizes mutual information of $y_x = f_x + \epsilon_x$ about f:
$x_n = \arg\max_{x\in S} I(f_A; y_x|D_{n-1})$
If $f\sim \text{GP}(\mu,k)$, then:
$I(f_A; y_x|D_{n-1}) = \frac12\log\left(\frac{\text{Var}[y_x|D_{n-1}]}{\text{Var}[y_x|f_A, D_{n-1}]}\right)$

Safe Bayesian Optimization
$x_n = \arg\max_{x\in\hat S_n = \{x | u_n^g(x)\ge 0\}} u_n^f(x)$

Batch Active Learning | ProbCover
$G=(X,E)$, $E=\{(x,x') \mid \|x-x'\|\le\delta\}$ $L\leftarrow\emptyset$ $\forall i = 1,2,\dots,b\{$ $\hat x \leftarrow \arg\max_{x\in X}|\{x' \mid (x,x')\in E, x'\in X\}|$ $L\leftarrow L\cup\{\hat x\}$ | $E\leftarrow E\setminus(\{\hat x\}\times(B_\delta(\hat x)\cap X))\}$

Max Mean Discrep. $MMD^2(\mathcal F, X, Y) = \sup_{\|f\|_{\mathcal H}\le1}[(\mathbb E_P[f(x)] - \mathbb E_q[f(y)])^2 = (\mathbb E_P\langle\phi(x), f\rangle_{\mathcal H} - \mathbb E_q\langle\phi(y),f\rangle_{\mathcal H})^2 = \langle\mu_x - \mu_y, f\rangle^2_{\mathcal H}] = \|\mu_x-\mu_y\|^2_{\mathcal H} = \mathbb E[k(x,x')] + \mathbb E[k(y,y')] - 2\mathbb E[k(x,y)]$

Convex Optimization

Given constrained optimization problem: $\min_{w\in\mathbb R^d} f(w)$: $g_{1:m}(w)=0, h_{1:n}(w)\le0$ it is convex if $f, g_{1:m}, h_{1:n}$ are convex and the feasible region is convex. The Lagrangian with Lagrange multipliers $\eta=(\lambda,\alpha)$: $L(\eta,w) = f(w)+\sum_{i\le m}\lambda_i g_i(w) + \sum_{j\le n}\alpha_j h_j(w)$ Any **optimal solution** W satisfies: $\nabla_w L(\eta, W)=0, g_i(W)=0, h_j(W)\le0, \alpha_j\ge0$ the **Dual Problem** is and satisfies $\forall w$: $\max_{\alpha\ge0,\lambda}[\theta(\eta) := \inf_w L(\eta,w)] \le f(w^*)$
strong duality if $\theta(\eta^*)=f(w^*)$
Slater's cond. if $\exists w_0$ feasible: $h_{1:n}(w_0)<0$ strong duality→ $w^\star$: $f(w^*)=L(\eta^*,w^*)$ and $\alpha_j h_j(w^*)=0$, $\forall j\le n$.

Support Vector Machine (SVM)

Convex constrained optimization problem with strong duality (if linearly separable). $\mathbf x_i$ support vectors, $y_i\in\{-1,+1\}$.
$\min_{w,w_0|\forall i\le n: y_i(w^\top x_i+w_0)\ge1}$ $\frac12\|w\|^2$
Lagrangian: $\mathcal L(w,w_0,\alpha) = \frac12\|w\|^2 + \sum_{i=1}^n\alpha_i(1-y_i(w^\top x_i+w_0))$ $\alpha_i\ge0$.
KKT: $w^*=\sum_{i=1}^n \alpha_i y_i x_i$ $\sum_{i=1}^n\alpha_i y_i = 0$
Dual: $\max_{\alpha\ge0:\sum_{i=1}^n\alpha_iy_i=0}$ $L(\alpha)$
$\cdot L(\alpha) = \sum_1^n\alpha_i - \frac12\sum_1^n\sum_{j=1}^n \alpha_i\alpha_j y_i y_j x_i^\top x_j$
The optimal hyperplane is given by
$\mathbf w^* = \sum_{i=1}^n \alpha_i^* y_i \mathbf x_i$
$w_0^* = -\frac12(\min_{y_i=1}\mathbf w^{*T}\mathbf x_i + \max_{y_i=-1}\mathbf w^{*T}\mathbf x_i)$
Only Support Vectors ($\alpha_i^*\ne0$) contribute.
Optimal Margin: $\mathbf w^T\mathbf w = \sum_{i\in SV}\alpha_i^*$
Discrim.: $g^*(\mathbf x) = \sum_{i\in SV} y_i\alpha_i^*\mathbf x_i^T\mathbf x_i + w_0^*$
class $= \text{sign}(\mathbf x^T\mathbf w^* + w_0^*)$

Soft Margin SVM
Introduce slack to relax constraints. C controls margin maximization vs. constraint violation.
$\min_{\xi_i\ge0,w,w_0|\forall i\le n: y_i(w^\top x_i+w_0)\ge1-\xi_i}$ $\frac12\|w\|^2 + C\sum_{i=1}^n\xi_i$
Lagrangian: $L(\mathbf w, w_0,\xi,\alpha,\beta) = \frac12\mathbf w^T\mathbf w + C\sum_{i=1}^n\xi_i - \sum_{i=1}^n\alpha_i[z_i(\mathbf w^T\mathbf y_i + w_0) - 1 + \xi_i] - \sum_{i=1}^n\beta_i\xi_i$
Dual Problem same as usual SVM but with supplementary constraint: $C\ge\alpha_i\ge0$
KTT Conditions: $\alpha_i^*(z_i(w^Ty_i+w_0)-1+\xi)=0, \xi_i(\alpha_i - C)=0$
You should solve α via quadratic optimisation. Optimal hyperplane and classification as normal SVM. Optimal slack: $\xi_i^* = \max(0, 1-y_i(w^{*T}x_i+w_0^*))$
$\xi_i^* = \mathcal L^{\text{hinge}}(y_i, w^{*T}x_i + w_0^*)$

Non-Linear SVM
Use kernel in discriminant function: $g(\mathbf x) = \sum_{i,j=1}^n \alpha_i z_i K(\mathbf x_i, \mathbf x)$
E.g solve the XOR Problem with: $K(x,y) = (1+x_1y_1+x_2y_2)^2$

Multiclass SVM

$\forall$class $y\in\{1,2,\cdots,M\}$ we introduce $\mathbf w_y$ and define our problem: (**w** is v-stacked)
$min_w \frac12 w^Tw = min_{\{w_y\}_{n=1}^M}\sum_{y=1}^M w_y^T w_y$
s.t. $(\mathbf w_{y_i}^T\mathbf x_i + w_{y_i,0}) - \max_{y\ne y_i}(\mathbf w_y^T\mathbf x_i + w_{y,0}) \ge 1, \forall\mathbf x_i\in\mathcal X$
classification: $\hat y = argmax_y(w_y^Tx + w_{y,0})$

Ensemble Methods

Combining Regressors

Set of estimators: $\hat f_1(x),\cdots,\hat f_B(x)$
simple average: $\hat f(x) = \frac1B\sum_{i=1}^B\hat f_i(x)$
$\text{Bias}[\hat f(x)] = \frac1B\sum_{i=1}^B \text{Bias}[f_i(x)]$
$\mathbb V[\hat f(x)]\approx\frac{\sigma^2}{B}$ if the estimators are uncorrelated.

Combining Classifiers

Input: classifiers $c_1(x),\cdots,c_B(x)$
Infer $\hat c_B(x) = \text{sgn}(\sum_{b=1}^B\alpha_b c_b(x))$
with weights $\{\alpha_b\}_{b=1}^B$
Requires diversity of the classifiers.

Bagging

Train on bootstrapped subsets. Covariance small, variance similar, bias weakly affected. **Random Forest** Collection of uncorr. decision trees. Partition data space recursively. Grow the tree sufficiently deep to reduce bias. (random sample cuts to reduce bias). Prediction with voting.
Boosting (Weak to avoid overfitting)
Combine uncorr. weak learners in sequence. Coeff. of $\hat c_{b+1}$ depend on $\hat c_b$'s results
AdaBoost (minimizes exp. loss)
Init: $\mathcal X = \{(x_1,y_1),\cdots,(x_n,y_n)\}, w_i^{(1)} = \frac1n$
Fit $\hat c_b(x)$ to $\mathcal X$ weighted by $w^{(b)}$
$\epsilon_b = \sum_{i=1}^n w_i^{(b)}\mathbb I_{\{\hat c_b(x_i)\ne y_i\}} / \sum_{i=1}^n w_i^{(b)}$
$\alpha_b = \log\frac{1-\epsilon_b}{\epsilon_b} > 0$
$w_i^{(b+1)} = w_i^{(b)}\exp(\alpha_b\mathbb I_{\{\hat c_b(x_i)\ne y_i\}})$
return $\hat c_B(x) = \text{sgn}(\sum_{b=1}^B\alpha_b\hat c_b(x))$
Best approx. at log-odds ratio.
Like stagewise-additive modeling.
Difference Boosting: identical $\mathcal D$, $\forall c(x)$ prediction weighted on accuracy, Bagging: varies $\mathcal D$, gives same importance. **Notes** AdaBoost gives high weight to hard-to-classify samples (maybe outliers). Bagging, if imbalanced dataset maybe $\mathcal Z$ missing a class. then, make the bootstrap size large enough s.t. at least one point is included.

Logistic Regression

$log\frac{P(y=1|x)}{P(y=-1|x)} = \sum_{b=1}^B c_b(x) =: F(x)$
$P(y=1|x) = \frac{exp(F(x))}{1+exp(F(x))}$

PAC learning

Exp./Gen. err: $\mathcal R(\hat c_n) = \mathbb P_{X,Y}(\hat c_n(x)\ne c(x))$
Emp. err.: $\hat{\mathcal R}_n(\hat c_n) = \frac1n\sum_{i=1}^n 1\{\hat c_n(x_i)\ne y_i\}$
Eff. PAC learnable: $\mathcal A$ can learn a concept class $\mathcal C$ from $\mathcal H$ if, given a sufficiently large sample, it outputs a hypothesis that generalizes well with high probability.
$0<\epsilon<\frac12, 0<\delta<\frac12, (X,Y)\in\mathcal X\times\{0,1\}$:
If $n\ge poly(\frac1\epsilon,\frac1\delta, dim(\mathcal X))$, then $\mathbb P_{X,Y}(\mathcal R(\hat c_n) - \inf_{c\in\mathcal C}\mathcal R(c)\le\epsilon)\ge1-\delta$.

VC Inequality

Select ERM. Under uniform convergence: $\mathbb P(\mathcal R(\hat c_m^*) - \inf_{c\in\mathcal C}\mathcal R(c)>\epsilon)\le$
$\mathbb P\left(\sup_{c\in\mathcal C}|\hat{\mathcal R}_n(c) - \mathcal R(c)| > \frac\epsilon2\right)$:
$P(\sup|\dots|>\epsilon)\le 2|\mathcal C|\exp(-2n\epsilon^2)$
$P(\sup|\dots|>\epsilon)\le 9n^{V_{\mathcal C}}\exp\left(-\frac{n\epsilon^2}{32}\right)$
$\mathbb P[\mathcal R(\hat c) - \inf_{c\in\mathcal C}\mathcal R(c)>\epsilon] < 1-\delta$.
Def R.H.S.$\le\delta$: $\epsilon = \sqrt{\frac{\log N - \log(\delta/2)}{2n}}$.
Consider $\mathcal H_\epsilon = \{h\in\mathcal H: R(h)>\epsilon\}$. We bound the probability of bad learning for consistent learn.: $P(\exists h\in\mathcal H_\epsilon : \hat R(h)=0)\le\sum_{h\in\mathcal H_\epsilon}P(\hat R(h)=0)$ $\le |\mathcal H_\epsilon|(1-\epsilon)^m \le |\mathcal H|\exp(-m\epsilon)\le\delta$ $\Rightarrow m\ge\frac1\epsilon\left(\log(|\mathcal H|)+\log\left(\frac1\delta\right)\right)$

VC dimension

classifier can shatter any n but no some $n+1$ points. **Examples:** $(-\infty,a]=1$ all intervals in R: $V_C=2$ For k intervals, $2k$ half planes in R^2: 3 for unit circles 3 convex polygons in R^2: ∞ convex polygons in R^2 with at most k vertices: $2k+1$

Nonparametric Bayesian methods

$\text{Beta}(x|a,b) = B(a,b)^{-1}x^{a-1}(1-x)^{b-1}$: prob. of Bernoulli proc. after observing $a-1$ success and $b-1$ failures. Multivariate case: Dirichlet distr. that will give multivar. probs, *based on finite* counts! But we don't know exactly which multivar. distribution works. With more data, we update the Dirichlet distribution. Is a conjugate prior.
Stick-breaking Dirichl. proc.
Repeatedly draw from $\text{Beta}(x|1,\alpha)$ with fixed α, but from reducing stick: $\rho_k = \beta_k(1-\sum_{i=1}^{k-1}\rho_i)$. The prior:
$\mathbb P[z_i=k|z_{-i},\alpha] = \begin{cases}\frac{N_{k,-i}}{\alpha+N-1} & \text{existing } k\\ \frac{\alpha}{\alpha+N-1} & \text{otherwise}\end{cases}$

Final Gibbs sampler:
$\mathbb P[z_i=k|z_{-i},\alpha,\mu] = \begin{cases}\frac{N_{k,-i}}{\alpha+N-1}p(x_i|x_{-i,k},\mu) & \text{existing } k\\ \frac{\alpha}{\alpha+N-1}p(x_i,\mu) & \text{otherwise}\end{cases}$

Gibbs sampling

Init: assign all data to a cluster, with prior π_i, with $\sum_{k=1}^K\pi_i<1$ (s.t. new clusters possible). E.g. with stick-breaking. Then remove x from k and compute new θ_k, then compute Gibbs sampler prob. (CRP), and sample the new cluster assignment $z_i\sim p(z_i|x_{-i},\theta_k)$. If cluster is empty, remove it and decrease K.