PROBABILITY THEORY

Product $p(x,y) = p(y|x)\,p(x)$

Bayes $p(x|y)\,p(y) = p(y|x)\,p(x)$

Marg. $p(x) = \int_x p(x,y)dy = \int_x p(x|y)\,p(y)dy$

Entropy $H(X) = -\int p(x)\log p(x)dx$

$H(X,Y) = H(X|Y) + H(Y) = H(Y|X) + H(X)$

$H(X|Y) = H(X)$ if $X \perp\!\!\!\perp Y$

$H(\mathcal{N}(M,\bar{\Sigma})) = 1/2 \ln|2\pi e\bar{\Sigma}|$

IG $I(X,Y) = H(X) - H(X|Y) = I(Y;X)$

↳ Monotone $F(A\cup\{x\}) - F(A) \geq$

Submodular $F(B\cup\{x\}) - F(B)$ $\forall A\subseteq B, \forall x$

KL-Div $KL(q\|p) = \int q(\theta)\log\frac{q(\theta)}{p(\theta)}d\theta$

↳ $KL(q\|p) \geq 0$ $KL(q\|p) \neq KL(p\|q)$

Properties

$\mathbb{E}[aX+bY] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y]$

$\mathrm{Var}[X] = \mathbb{E}[(x-\mathbb{E}[x])(x-\mathbb{E}[x])^T]$
$= \mathbb{E}[x^2] - \mathbb{E}[x]^2$

$\mathrm{Cov}(x,y) = \mathbb{E}[(x-\mathbb{E}[x])(y-\mathbb{E}[y])^T]$
$= \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]$

$\mathbb{V}[aX \pm bY] = a^2\mathbb{V}[x] + b^2\mathbb{V}[x] \pm 2ab\,\mathrm{Cov}(x,y)$

Joint Expectation $\mathbb{E}_{x,y}[\cdot] = \mathbb{E}_x[\mathbb{E}_y[\cdot|x]]$

Tower Rule $\mathbb{E}_x[x] = \mathbb{E}_y[\mathbb{E}_{x|y}[x|y]]$

Law of Unc. Sta. $\mathbb{E}_x[g(x)] = \int g(x)\,p(x)\,dx$

Tot. Var. $\mathbb{V}[y] = \mathbb{E}_x[\mathbb{V}[y|x]] + \mathbb{V}[\mathbb{E}_x[y|x]]$

Jensen $g(\mathbb{E}[x]) \leq \mathbb{E}[g(x)]$ $g(\cdot)$ convex

Cauchy-Schwarz $|\mathbb{E}[x,y]|^2 \leq \mathbb{E}[x^2]\,\mathbb{E}[y^2]$

$\mathbb{E}[x^T] = \mathbb{E}[x]^T$ $\mathbb{V}[Ax] = A\,\mathbb{V}[x]\,A^T$

MultiVar Gaussian $MX \sim \mathcal{N}(M\mu_x, M\Sigma M^T)$

Sum: $x \perp\!\!\!\perp y \Rightarrow \mathcal{N}(\mu_x+\mu_y, \Sigma_x+\Sigma_y)$

Cond. $x = \begin{bmatrix}x_1\\x_2\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix}, \begin{bmatrix}\bar{\Sigma}_{11} & \bar{\Sigma}_{12}\\ \bar{\Sigma}_{21} & \bar{\Sigma}_{22}\end{bmatrix}\right)$

↳ $x_1|x_2 \sim \mathcal{N}(\mu_{x1|x2}, \bar{\Sigma}_{x1|x2})$

↳ $\mu_{x1|x2} = \mu_1 + \bar{\Sigma}_{12}\bar{\Sigma}_{22}^{-1}(x_2-\mu_2)$

↳ $\bar{\Sigma}_{x1}|x_2 = \bar{\Sigma}_{11} - \bar{\Sigma}_{12}\bar{\Sigma}_{22}^{-1}\bar{\Sigma}_{21}$

PDF $\frac{1}{\sqrt{2\pi^n|\Sigma|}}\exp\left(-\frac{1}{2}(x-\mu)^T\bar{\Sigma}^{-1}(x-\mu)\right)$

BAYESIAN LINEAR REGRESSION

Model $y = w^T\phi(x) + \varepsilon$ $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$

Likelihood $p(y|x,w) = \mathcal{N}(w^Tx, \sigma_n^2)$

Prior $p(w) = \mathcal{N}(0, \sigma_p^2 I)$

Inference $p(w|x,y) = \mathcal{N}(\bar{\mu}, \bar{\Sigma})$

↳ $\bar{\mu} = (X^TX + \frac{\sigma_n^2}{\sigma_p^2}I)^{-1}X^Ty$

↳ $\bar{\Sigma} = (\sigma_n^{-2}X^TX + \sigma_p^{-2}I)^{-1}$

$w_{MLE} = \arg\max_w p(y|x,w) = (X^TX)^{-1}X^Ty$

$w_{MAP} = \arg\max_w p(w|x,y) = \bar{\mu}$

Predict $\mathcal{O}(Nd^2)$

$p(y^*|x,y,x^*) = \mathcal{N}(\bar{\mu}^Tx^*, \underbrace{x^{*T}\bar{\Sigma}x^*}_{1} + \underbrace{\sigma_n^2}_{2}\cdot I)$

1. Epistemic Uncert. Lack of Data
2. Aleatoric Uncert. Irred Noise Same x_t Diff y_t

Hyperparams 1. $\hat{\lambda} = \sigma_n^2/\sigma_p^2$ by Ridge Regr.

2. $\hat{\sigma}_n^2 = \frac{1}{n}\sum_i(y_i - \hat{w}^Tx_i)^2$ MAP

3. Solve $\hat{\sigma}_p^2 = \hat{\sigma}_n^2/\hat{\lambda}$

GAUSSIAN PROCESS $\mathcal{O}(N^3)$

Model $y = w^T\phi(x) + \varepsilon$ $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

Prior $p(w) = \mathcal{N}(0,1)$

$f \sim GP(M,K)$ $f = \begin{bmatrix}f(-\infty)\\ f(+\infty)\end{bmatrix} \in \mathbb{R}^\infty$

$\begin{bmatrix}f(x_1)\\ \vdots \\ f(x_N)\end{bmatrix} \in \mathcal{N}(\mu, K)$ $\mu = \begin{bmatrix}M(x_1)\\ \vdots \\ \mu(x_N)\end{bmatrix}$

$K = \begin{bmatrix}K(x_1,x_1) & \cdots & K(x_1,x_N)\\ \vdots & & \vdots \\ K(x_n,x_1) & \cdots & K(x_n,x_N)\end{bmatrix}$ $K_{x,A} = \begin{bmatrix}K(x,x_1)\\ \vdots \\ K(x,x_N)\end{bmatrix}$

PREDICT $p(f|X_{1:n}, y_{1:n}) = GP(M', k')$

$M'(x) = M(x) + K_{x,A}(K_{AA} + \sigma_n^2 I)^{-1}(y_A - \mu_A)$

$K'(x,x')$ $A = \{x_1, \ldots, x_n\}$ observed

$= K(x,x') - K_{x,A}(K_{AA} + \sigma_n^2 I)^{-1}K_{x',A}^T$

Hyperparams $\hat{\theta} = \arg\max_\theta p(y_{1:n}|x_{1:n},\theta)$

$= \arg\max_\theta \int_f p(y_{1:n}|f, x_{1:n}, \theta)p(f|\theta)df$

Kernels $K(x,x') = \phi(x)^T\phi(x')$

$K(x,x') = K(x',x)$ K_{AA} PSD $\forall A \subseteq \mathcal{X}$

$K(x,x') = \mathrm{Cov}[f(x), f(x')]$ Symmetric

$\begin{bmatrix}a & b\\ c & d\end{bmatrix}^{-1} = \frac{1}{ad-bc}\begin{bmatrix}d & -b\\ -c & a\end{bmatrix}$ Positive definite Valid

$K(x_1,x_2) = \mathrm{Cov}(f(x_1), f(x_2))$

$= \mathbb{E}[(f(x_1) - M(x_1))\cdot(f(x_2) - M(x_2))]$

STATIONARY $K(x,x') = K(x-x')$

ISOTROPIC $K(x,x') = k(\|x-x'\|_2)$

KALMAN FILTER $Y_t \perp\!\!\!\perp X_{1:t-1} | x_t$ or $x_{t-1} \perp\!\!\!\perp X_{1:t-1}, Y_{1:t-1} | x_t$

Motion $P(x_{t+1}|x_t) = \mathcal{N}(x_{t+1}; Fx_t, \bar{\Sigma}_x)$

$x_{t+1} = Fx_t + \varepsilon_t$ $\varepsilon_t \sim \mathcal{N}(0, \bar{\Sigma}_x)$

Sensor $P(y_{t+1}|x_t) = \mathcal{N}(y_t; Hx_t, \bar{\Sigma}_y)$

$y_t = Hx_t + \eta_t$ $\eta_t \sim \mathcal{N}(0, \bar{\Sigma}_y)$

Update $\mu_{t+1} = F\mu_t + K_{t+1}(y_{t+1} - HF\mu_t)$

$\Sigma_{t+1} = (I - K_{t+1}H)(F\bar{\Sigma}_tF^T + \bar{\Sigma}_x)$

Gain $K_{t+1} = (F\bar{\Sigma}_tF^T + \bar{\Sigma}_x)\cdot[H^T(H(F\bar{\Sigma}_tF^T + \bar{\Sigma}_x)H^T + \bar{\Sigma}_y)^{-1}$

Bayesian Filtering Keep track X_t. $P(X_t) = \mathcal{N}(M, \bar{\Sigma})$ assume $P(x_t|y_{1:t-1})$

Predict $p(x_{t+1}|y_{1:t}) = \int p(x_{t+1}|x_t)p(x_t|y_{1:t})dx_t$

BAYESIAN LOGISTIC REGRESSION

Model $w = \mathrm{sign}(w^Tx)$

Likelihood $p(y|x,w) = \mathrm{Bern}(\sigma(yw^Tx))$

$= \prod_{i=1}^n p(y_i|x_i;w) = \prod_{i=1}^n \sigma(y_i \cdots w^Tx_i)$

Prior $p(w) = \mathcal{N}(0, \sigma_p^2 I)$

Laplace Approx $p(w|x,y) \approx \mathcal{N}(\hat{w}, \Lambda^{-1})$

$\hat{w} = \arg\min_w \frac{1}{2\sigma_p^2}\|w\|_2^2$ ← λ ↓ logistic loss $+\sum_{i=1}^n \log(1 + \exp(-y_iw^Tx_i))$

$\Lambda = -\nabla\nabla\log p(w|x_{1:n}, y_{1:n})$

$= \sum_{i=1}^n x_ix_i^T\pi_i(1-\pi_i)$

$= X^T\mathrm{diag}([\pi_i(1-\pi_i)]_i)X$

$\pi_i = \sigma(\hat{w}^Tx_i)$ ⚠ does not depend on $\bar{\Sigma}$

Predictions $p(y^*|x^*, x_{1:n}, y_{1:n})$

$\approx \int\sigma(y^*w^Tx)\mathcal{N}(w; \hat{w}, \Lambda^{-1})dw$

$= \int\sigma(y^*f)\mathcal{N}(f; \hat{w}^Tx^*, x^{*T}\Lambda^{-1}x^*)df$

VARIATIONAL INFERENCE

Approximates $p(\theta|x,y) \approx q_\lambda(\theta)$

where $q_\lambda(\theta) = \mathcal{N}(\mu = \hat{\theta}, \bar{\Sigma} = \Lambda^{-1})$

↳ $\hat{\theta} = \arg\max_\theta p(\theta|x,y)$

$\Lambda = -\nabla\nabla\log p(\theta|x,y)$ Hessian

Derivation MAP $\psi(\theta) = \log p(\theta|x_{1:n}, y_{1:n})$

$\hat{\psi}(\theta) = \psi(\hat{\theta}) + (\theta-\hat{\theta})^T\nabla\psi(\hat{\theta}) + \frac{1}{2}(\theta-\hat{\theta})^T\nabla\nabla\psi(\hat{\theta})(\theta-\hat{\theta})$

$= \psi(\hat{\theta}) + \frac{1}{2}(\theta-\hat{\theta})^TH_\psi(\hat{\theta})(\theta-\hat{\theta})$

Forward KL $KL(p\|q)$ covers full pdf

Reverse KL $KL(q\|p)$ focuses on $\mathrm{mode}_{\max}$

VI $q^* \in \arg\min_q KL(q\|p)$

$= \arg\max_{q\in\mathcal{Q}} \mathbb{E}_{\theta\sim q}[\log p(y|\theta)] - KL(q\|p)$

$= \arg\max_{q\in\mathcal{Q}} \mathbb{E}_{\theta\sim q}[\log p(y,\theta)] + H(q)$ ELBO

Reparam. Trick as q dep. on var. params

$\nabla\mathbb{E}_{\theta\sim q_\lambda}[f(\theta)] = \nabla_\lambda\mathbb{E}_{\varepsilon\sim p}[f(g(\varepsilon,\lambda))]$

distr. dep. on λ new distr. $g(\varepsilon,\lambda) = \theta$

e.g. Gaussians $q(\theta|\lambda) = \mathcal{N}(\theta; M, \bar{\Sigma})$

$\lambda = [M, C]$ $\bar{\Sigma} = CC^T$ $\theta = C\varepsilon + M$

$\phi(\varepsilon) = \mathcal{N}(0,I)$ $\varepsilon = C^{-1}(\theta - M)$ $\phi(\varepsilon) = q(\theta|\lambda)|C|$

MARKOV-CHAIN MONTE-CARLO

Hoeffding's Inequality

$P(|\mathbb{E}_p[f(x)] - \frac{1}{N}\sum_{i=1}^N f(x_i)| > \varepsilon) \leq 2\exp(-2N\varepsilon^2/C^2)$

Ergodic MC if $\exists$ a finite t s.t. every state can be reached from every state in exactly t steps.

Stationary Distributions Ergodic MC has unique and positive $\pi(x) > 0$ s.t. $\forall x$ $\lim_{N\to\infty} p(X_N = x) = \pi(x)$

Indep of $p(x_1)$ All MC made ergodic $P_{ij} > 0$

Detailed Balance Eq. ↑ $\pi = \pi P$, $\sum\pi_i = 1$

$Q(x)P(x'|x) = Q(x')P(x|x')$

Ergodic Theorem if MC Ergodic

$\Rightarrow \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^N f(x_i) = \sum_{x\in D}\pi(x)f(x)$

$\mathbb{E}[f(x)]$ via MCMC

$\approx (T-t_0)^{-1}\sum_{T=t_0+1}^T f(x^{(T)})$

Metropolis Hastings

Proposal $x' \sim R(x'|x)$, Given $X_t \sim x$

Acceptance: with prob. $\alpha = \min\left\{1, \frac{Q(x')}{Q(x)}\frac{R(x|x')}{R(x'|x)}\right\}$

→ Set $X_{t+1} = x'$ $Q(x) = \exp(-f(x))$

Otherwise → Prob $1-\alpha$ set $X_{t+1} = x$

Metropolis Adjusted Langevin Algorithm

Take ∇ info into account to prefer proposals into Region with high density

$R(x'|x) = \mathcal{N}(x'; x - \tau\nabla f(x); 2\tau I)$

+ Log-concave → MALA Converges to $\pi(x)$

- Requires full energy function f ↴

S.G. Langevin Dynamics

SGD + Gaussian Noise, gurantee to converge if $\eta_t \in \Theta(t^{-1/3})$

Gibbs Sampling

$X^{[t]} = [x_1^{(t)}, \ldots, x_n^{(t)}]^T$

for $i \in \{1, 2, \ldots, n\}$:

$x_i^{(t+1)} \sim p(x_i|\, x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \ldots, x_n^{(t)})$

BAYESIAN DEEP LEARNING

Homoscedastic: ε Heteroscedastic: $\varepsilon(x)$

Prior $p(\theta) = \mathcal{N}(\theta; 0, \sigma_p^2 I)$ on weights

Likelihood $p(y|x,\theta) = \mathcal{N}(y; f_\mu(x,\theta), \exp(f_{\sigma^2}(x,\theta))$

MAP

$\hat{\theta} = \arg\min_\theta -\log p(\theta) - \sum_{i=1}^n \log p(y_i|x_i,\theta)$

$= \arg\min_\theta \lambda\|\theta\|_2^2 + \frac{1}{2}\sum_{i=1}^n [\sigma(x_i;\theta)^{-2}\cdot\|y_i - \mu(x_i;\theta)\|^2 + \log\sigma(x_i;\theta)^2]$

Predict $p(y^*|x,y,x^*) = \int p(y^*|x^*\theta)\cdot p(\theta|x,y)d\theta$ Intractable $\theta \in \mathbb{R}^{n>>0}$

VI Given Variational posterior q Can approx predictive distributions by Sampling

$p(y^*|x^*, x_{1:n}, y_{1:n})$

$\approx \mathbb{E}_{\theta\sim q(\cdot|\lambda)}[p(y^*|x^*,\theta)]$

$\approx 1/m\sum_{j=1}^m p(y^*|x^*, \theta^{(j)}) = \mu(x^*)$

$\mathbb{V}[y^*|x,y,x^*] =$
$\mathbb{E}[\mathbb{V}[y^*|x^*,\theta]] + \mathbb{V}[\mathbb{E}_\theta[y^*|x^*,\theta]]$
Aleotoric Uncert. Epistemic Uncert.
$\approx \frac{1}{m}\sum_{j=1}^m \sigma^2(x^*,\theta^{(j)}) +$
$\frac{1}{m}\sum_{j=1}^m (\mu(x^*,\theta^{(j)}) - \bar{\mu}(x^*))^2$

Dropout as VI $q(\theta|\lambda) = \pi_j q_j(\theta_j|\lambda_j)$
$q_j(\theta_j|\lambda_j) = p\delta_0(\theta_j) + (1-p)\delta_{\lambda_j}(\theta_j)$

MCMC for Neural Networks

1. **Subsampling** $\approx 1/T \sum_{j=1}^T p(y^*|x^*,\theta^{(j)})$

2. **Gaussian Approximation**
Keep track of a Gaussian Approx of the params $q(\theta|\mu_{1:d}, \sigma^2_{1:d})$
$\mu_i = 1/T \sum_{j=1}^T \theta_i^{(j)}$, $\sigma_i^2 = 1/T \sum_{j=1}^T (\theta_i^{(j)} - \mu_i)^2$
$\theta^{(j)} \in \mathbb{R}^d$ sampled after burn-in

3. **Probabilistic Ensembles for NNs**
Train m NNs on bootstrap datasets
$p(y^*|x^*,\theta) \approx 1/m \sum_{j=1}^m p(y^*|x^*,\theta^{(j)})$

ECE $\sum_{m=1}^M \frac{|B_m|}{n} |acc(B_m) - conf(B_m)|$
$acc(B_m) = |B_m|^{-1} \sum_{i \in B_m} \mathbb{I}\{\hat{y}_i = y_i\}$
$conf(B_m) = |B_m|^{-1} \sum_{i \in B_m} \hat{p}_i$

ACTIVE LEARNING

Uncertainty Sampling Pure Exploitative
$x_{t+1} = \arg\max_{x \in \mathcal{X}} \Delta_I(x|s_t)$
$= \arg\max_{x \in \mathcal{X}} IG(f_x; y_x | y_{s_t})$
$= \arg\max_{x \in \mathcal{X}} \frac{1}{2}\log(1 + \sigma_t^2(x)/\sigma_n^2)$
$= \arg\max_{x \in \mathcal{X}} \sigma_t^2(x)$ (Const.)
$\Delta_I(x|s_t) = F(s_t \cup \{x\}) - F(s_t)$

Greedy Algorithm provides Const. Approx.
$F(S_T) \geq (1 - \frac{1}{e}) \max_{S \subseteq D, |S| \leq T} F(S)$

- Fails to distinguish Epistemic/Aleatoric
- Heteroscedastic Case, most uncertain outcomes ≠ most informative

Maximizing IG yields
$x_{t+1} \in \arg\max_x \frac{\sigma_f^2(x)}{\sigma_n^2(x)}$ epistemic / Aleatoric

Bayesian Active Learning By Disagreement
Classif. $x_{t+1} = \arg\max_{x \in \mathcal{X}} H[y_x | x_{1:t}, y_{1:t}]$
$= \arg\max_{x \in \mathcal{X}} H[y_x|x_{1:t}, y_{1:t}] -$
$\mathbb{E}_{\theta|x_{1:t},y_{1:t}} H[y_x|\theta]$

BAYESIAN OPTIMIZATION

Cumulative Regret
$R_T = \sum_{t=1}^T (\max_x f(x) - f(x_t))$
Sublinear $R_T/T \to 0$, implies $\max_t f(x_t) \to f(x^*)$

Algo $f \sim GP(\mu_0, k_0)$
for t=1 to T do:
$x_t \leftarrow \arg\max_{x \in \mathcal{X}} F(x; \mu_{t-1}, k_{t-1})$
$y_t \leftarrow f(x_t) + \varepsilon_t$
Bayesian update μ_t and k_t (NOT CONVEX)

UCB $(x; \mu, \sigma) = \mu(x) + \beta_t \sigma(x)$
β_t: Controls Exploration $\sigma(x) = \sqrt{k(x,x)}$

GP-UCB $R_T/T = O(\sqrt{\gamma_T/T})$
$\gamma_T = \max_{S \subseteq X, |S|=T} I(f; y_S)$

Bounds on IG due to submodularity
Linear: $\gamma_T = O(d \log T)$
Gaussian: $\gamma_T = O((\log T)^{d+1})$
Matern $\nu > \frac{1}{2}$: $\gamma_T = O(T^{\frac{d}{2\nu+d}} (\log T)^{\frac{2\nu}{2\nu+d}})$

Guarantees sublinearity / Convergence
$PI(x; \mu, \sigma) = P(f(x) > f^+)$
$= \Phi\left(\frac{\mu(x) - f^+}{\sigma(x)}\right)$ $Z(\cdot)_+ = \max(0, \cdot)$
$EI(x; \mu, \sigma) = \mathbb{E}_{f|x_{1:t}, y_{1:t}}[(f(x) - f^+)_+]$
$= (\mu(x) - f^+)\Phi(Z) + \sigma(x)\phi(Z)$

Thompson Sampling
Sample function $\tilde{f} \sim p(f|x_{1:t}, y_{1:t})$
Select $x_{t+1} \in \arg\max_{x \in D} \tilde{f}(x)$
Randomness of $\tilde{f}$ sufficient to trade exploration/exploitation

RL BASICS

Value Function $V^\pi(x) = r(x, \pi(x)) + \gamma \sum_{x'} p(x'|x, \pi(x)) V^\pi(x')$

Stochastic $V^\pi(x) = \sum_a \pi(a|x) (r(x,a) + \gamma \sum_{x'} p(x'|x,a) V^\pi(x'))$

Action-Value (Q) $Q^\pi(x,a) = r(x,a) + \gamma \sum_{x'} p(x'|x,a) V^\pi(x')$

Stochastic $Q^\pi(x,a) = r(x,a) + \gamma \sum_{x'} p(x'|x,a) \sum_{a'} \pi(a'|x) q^\pi(x',a')$

Expected Value $J(\pi) = \sum_{m=0}^\infty \gamma^m R_{t+m}$ (↑ expected total reward)

Multiple Trajectories
$J(\pi) = \sum_{i=1}^{|T|} P(T_i) J(T_i)$

$V^\pi(x)$ Closed form
$v^\pi = [V^\pi(1) \cdots V^\pi(n)]^T$
$r^\pi = [r(1, \pi(1)) \cdots r(n, \pi(n))]^T$
$p^\pi = \begin{bmatrix} p(1|1,\pi(1)) & \cdots & p(n|1,\pi(1)) \\ \vdots & \ddots & \vdots \\ p(1|n,\pi(n)) & \cdots & p(n|n,\pi(n)) \end{bmatrix}$
Each Row Sums to 1
$V^\pi = r^\pi + \gamma P^\pi V^\pi \Rightarrow V^\pi = (I - \gamma P^\pi)^{-1} R^\pi$
Invertible if $\gamma \in [0,1)$

Advantage Function $A^\pi(x,a) = Q^\pi(x,a) - V^\pi(x)$

Optimality $V^*(x) = \max_{\pi \in \Pi} V^\pi(x)$
$= \max_{a \in A} Q^*(x,a)$, $Q^*(x,a) = \max_{\pi \in \Pi} Q^\pi(x,a)$
one possible (stationary + deterministic) $\pi^*(x) = \arg\max_{a \in A} Q^*(x,a)$
V^* Unique π^* not always
↳ For infinite Horizon MDP, there exists optimal stationary + deterministic Policy.

POMDP can't observe x_t directly, only observation y_t → new belief states $b: \{1, \ldots, n\} \to [0,1]$, $\sum_x b(x) = 1$
Actions same as original MDP.
Transition Model: 1. Stochastic
observation $P(y_{t+1}=y|b_t, a_t) =$
$\sum_{x,x'} b_t(x) P(x'|x, a_t) P(y|x')$
2. State Update: $b_{t+1}(x') = 1/z \sum_x$
$b_t(x) P(x_{t+1}=x'|x_t=x, a) P(y_{t+1}|x')$
Reward: $r(b_t, a_t) = \sum_x b_t(x) r(x, a_t)$

Num Det Policies $\prod_{i=1}^S A_i$

MODEL-BASED RL

Value Iter. $V_{k+1}(x) = \max_a \{Q_k\}$

ε-optimal in polynomial # iterations ($\gamma < 1$)

Policy Iter. Initialize π_0
Until Convergence:
  Compute $U^\pi(x)$ $\forall x$
  Compute Greedy Policy π_G w.r.t V^π
  Set $\pi \leftarrow \pi_G$

Finds exact solution in polynomial # iterations! $\pi^* \in O^*(n^2 m/(1-\gamma))$
Guaranteed to monotonically improve

Rmax Implicit Exploration (Agent always tries to explore)
$r(x,a) = R_{max}$, $P(x^*|x,a) = 1$ $\forall x,a$
Repeat: Update $r(x,a)$, $P(x'|x,a)$
If observed 'enough', Recompute π according to P, r. (ε-optimal π in # steps polynomial in $|x||A|, 1/\varepsilon, \log(1/\delta)$ and R_{max}.

MODEL-FREE RL

TD on-policy follow π to obtain (x,a,r,x'), Update via bootstrapping
$\hat{V}^\pi(x) \leftarrow (1-\alpha_t)\hat{V}^\pi(x) + \alpha_t(r + \gamma \hat{V}^\pi(x'))$
α_t satisfies $\sum_t \alpha_t = \infty$ $\sum_t \alpha_t^2 < \infty$
∧ all x chosen ∞ often $\hat{V}^\pi \to V^\pi$
⊕ Low Variance ⊖ biased

Q-Learning off-policy
Estimate $\hat{Q}^*(x,a)$ observe x, a, x', r
$Q(x,a) \leftarrow (1-\alpha_t)Q(x,a) + \alpha_t(r + \gamma \max_a Q(x',a'))$ Same converge Conditions as TD

Optimistic Q Initialize $\hat{Q}^*(x,a) = R_{max}/1-\gamma \prod_{t=1}^{T_{init}} (1-\alpha_t)^{-1}$
At time t, pick $a_t \in \arg\max_a \hat{Q}^*(x_t,a)$

Convergence same as Rmax
Q-learning: Mem $O(nm)$, RT: $O(m)/\varepsilon$

FUNCTION APPROX RL

$Q(x,a;\theta) = \theta^T \phi(x,a)$
$L(\theta) = \frac{1}{2}(\hat{Q}^*(x,a;\theta) - r - \max_{a'} \hat{Q}^*(x',a';\theta_{old}))^2$
Until Converged
  In state x, pick a
  Observe x', reward r
  $\theta \leftarrow \theta - \alpha_t \delta \nabla_\theta Q(x,a;\theta)$
$\delta := Q(x,a;\theta) - r - \gamma \max_{a'} Q(x',a';\theta)$

Deep Q-Learning (DQN)
$\ell_{DQN}(\theta) = \frac{1}{2} \sum_{(x,a,r,x') \in D} (r + \gamma \max_{a'} Q(x',a';\theta^{old}) - Q(x,a;\theta))^2$
Suffers „Maximation bias" ↓

DDQN Current network for argmax
$\ell_{DDQN}(\theta) = \frac{1}{2} \sum_{(x,a,r,x') \in D} (r + \gamma Q^*(x', a^*(x';\theta); \theta^{old}) - Q^*(x,a;\theta))^2$
where $a^*(x';\theta) = \arg\max_{a' \in A} Q^*(x',a';\theta)$

POLICY GRADIENT (↙ Unbiased but high Variance)
REINFORCE $\theta \leftarrow \theta + \eta \gamma^t G_t \nabla_\theta \log \pi(A_t|X_t;\theta)$
Reward-To-Go $G_t = \sum_{t=t}^T \gamma^{t-t} r_t$

Online Actor Critic
At time t after observing (x,a,r,x')
$\theta_\pi \leftarrow \theta_\pi + \eta_t Q(x,a;\theta_Q) \nabla \log \pi(a|x;\theta_\pi)$
$\theta_Q \leftarrow \theta_Q - \eta_t (Q(x,a;\theta_Q) - r - \gamma Q(x', \pi(x';\theta_\pi);\theta_Q)) \nabla Q(x,a;\theta_Q)$
Under Compatibility Conditions Guaranteed to Improve. Baselines to reduce Var