1. Probability & Info Theory

MLE, MAP: $\hat\theta_{\text{MLE}} = \arg\max_\theta \prod_i p(y_i|x_i,\theta) = \arg\max_\theta \sum_i \log p(y_i|x_i,\theta)$ **GaussianNoiseLinReg**: $y=\theta^\top x+\varepsilon$, $\varepsilon\sim\mathcal N(0,\sigma^2)$ MLE ≡ 最小二乘: $\hat\theta_{\text{MLE}} = \arg\min\sum_i (y_i-\theta^\top x_i)^2 = (X^\top X)^{-1}X^\top y$
$\hat\theta_{\text{MAP}} = \arg\max_\theta p(\theta|D) = \arg\max_\theta p(D|\theta)p(\theta) = \arg\min_\theta \underbrace{-\log p(\theta)}_{正则} + \underbrace{-\log p(D|\theta)}_{拟合}$
Prior→Regularizer: Gaussian $\mathcal N(0,\sigma_p^2 I)\to$ L2: $\lambda\|\theta\|^2$, $\lambda=\frac{1}{\sigma_p^2}$; Laplace$(0,b)$ → L1: $\lambda\|\theta\|_1$, $\lambda=\frac1b$

Posterior ∝ Likelihood × Prior: $p(\theta|D)\propto p(D|\theta)\cdot p(\theta)$, log 之, $\log p(\theta|D)=\log p(D|\theta)+\log p(\theta)-\log p(D)$, where $\log p(D)$"const w.r.t" θ.

Prob: **Product**: $\mathbb P(X_{1:n}) = \mathbb P(X_1)\prod_{i=2}^n \mathbb P(X_i|X_{1:i-1})$ **Bayes**: $\mathbb P(X|Y)=\frac{\mathbb P(Y|X)\mathbb P(X)}{\mathbb P(Y)}$ **Cond Indep**: $X\perp Y|Z \leftrightarrow \mathbb P(X,Y|Z)=\mathbb P(X|Z)\mathbb P(Y|Z)$ **Tower**: $\mathbb E_Y\left[\mathbb E_{X[X|Y]}\right]=\mathbb E[X]$ **LOTV**: $\mathbb V[X]=\mathbb E[\mathbb V[X|Y]]+\mathbb V[\mathbb E[X|Y]]$

Gaussian 性: $\mathcal N(x;\mu,\Sigma)=\frac{\exp(-\frac12(x-\mu)^\top\Sigma^{-1}(x-\mu))}{\sqrt{(2\pi)^d|\Sigma|}}$
Marginal: $X_A\sim\mathcal N(\mu_A,\Sigma_{AA})$ **Conditional**: $X_A|X_B\sim\mathcal N(\mu_{A|B},\Sigma_{A|B})$ $\mu_{A|B}=\mu_A+\Sigma_{AB}\Sigma_{BB}^{-1}(x_B-\mu_B)$; $\Sigma_{A|B}=\Sigma_{AA}-\Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}$
Linear: $Y=MX\sim\mathcal N(M\mu, M\Sigma M^\top)$
Sum: indep $X+X'\sim\mathcal N(\mu+\mu',\Sigma+\Sigma')$
Product (两高斯相乘): $\mathcal N(\mu_1,\Sigma_1)\cdot\mathcal N(\mu_2,\Sigma_2)\propto\mathcal N(\mu,\Sigma)$ $\Sigma^{-1}=\Sigma_1^{-1}+\Sigma_2^{-1}$ (precision 相加!) $\mu=\Sigma(\Sigma_1^{-1}\mu_1+\Sigma_2^{-1}\mu_2)$ **Precision Matrix**: $\Lambda:=\Sigma^{-1}$ (inverse covariance) 对角项: 条件方差倒数; 非对角: 条件相关性

Var & Cov: $\mathbb V[X]=\mathbb E[(X-\mathbb E[X])^2]$; $\text{Cov}[X,Y]=\mathbb E[XY]-\mathbb E[X]\mathbb E[Y]$ $\mathbb V[aX+bY]=a^2\mathbb V[X]+b^2\mathbb V[Y]+2ab\text{Cov}[X,Y]$ $\mathbb V[aX-bY]=a^2\mathbb V[X]+b^2\mathbb V[Y]-2ab\text{Cov}[X,Y]$ $\mathbb V[X+Y]=\mathbb V[X]+\mathbb V[Y]+2\text{Cov}[X,Y]$ $\mathbb V[X-Y]=\mathbb V[X]+\mathbb V[Y]-2\text{Cov}[X,Y]$

Info Theory: **Entropy**: $H[p]=-\mathbb E_p[\log p(x)]$; Gauss: $H=\frac12\log((2\pi e)^d|\Sigma|)$ **KL**: $\text{KL}(p\|q)=\mathbb E_p\left[\log\frac pq\right]\ge0$; Need $\text{supp}(q)\subseteq\text{supp}(p)$ Forward KL(p‖q): mean-seeking; Reverse KL(q‖p): mode-seeking **MI**: $I(X;Y)=H[X]-H[X|Y]=H[Y]-H[Y|X]\ge0$ **Info Gain 公式**: $I(X;Y,Z)=I(X;Y)+I(X;Z|Y)$ $I(X;Y,Z)-I(X;Y)=H[X|Y]-H[X|Y,Z]$ (条件减少熵!) **Info Never Hurts**: $I(X;Y)\ge0$ and $H[X|Y]\le H[X]$ 观测Y不会增加X的不确定性 **Cond MI**: $I(X;Y|Z)=H[X|Z]-H[X|Y,Z]$ **Gauss MI**: $I[X;Y]=\frac12\log\det(I+\sigma_n^{-2}\Sigma)$ for $Y=X+\varepsilon$

2. BLR: Linear Kernel GP

Model: $y=w^\top x+\varepsilon$, $\varepsilon\sim\mathcal N(0,\sigma_n^2)$ **Prior**: $w\sim\mathcal N(0,\sigma_p^2 I)$ (L2 正则/weight decay) **Posterior**: $w|X,y\sim\mathcal N(\mu,\Sigma)$ $\Sigma^{-1}=\sigma_n^{-2}X^\top X+\sigma_p^{-2}I$ (只依赖X!) $\mu=\sigma_n^{-2}\Sigma X^\top y$

Prediction: $x_*^\top\Sigma x_*$"epistemic"; σ_n^2"aleatoric": $y_*|x_*,X,y\sim\mathcal N(x_*^\top\mu, x_*^\top\Sigma x_*)+\sigma_n^2)$; **MAP=Ridge**: $\hat w=(X^\top X+\lambda I)^{-1}X^\top y$, $\lambda=\frac{\sigma_n^2}{\sigma_p^2}$ **BLogR**: Logistic Regression 无闭式解 (非高斯likelihood) 需 VI/Laplace/MCMC 近似 posterior

Online Update (Woodbury): $(A+xx^\top)^{-1}=A^{-1}-\frac{A^{-1}xx^\top A^{-1}}{1+x^\top A^{-1}x}$ $O(d^2)$ New data(x,y): $\Sigma_{t+1}=\Sigma_t-\frac{\Sigma_t xx^\top\Sigma_t}{1+x^\top\Sigma_t x}$ $\mu_{t+1}=\Sigma_{t+1}(\Sigma_t^{-1}\mu_t+yx)$

3. Gaussian Processes

Def: $y_i=f(x_i)+\varepsilon_i$, noise $\varepsilon_i\sim\mathcal N(0,\sigma_n^2)$. **Prior**: $f\sim\mathcal{GP}(\mu(x),k(x,x'))$. Finite set $A=\{x_1,...,x_m\}$, the vector $f(A)$多维 Gaussian, $f(X)\sim\mathcal N(\mu(A),K_{AA})$, $[K_{AA}]_{ij}=k(x_i,x_j)\in\mathbb R^{m\times m}$. $k(x_i,x_i)$:each points 自由度/方差; $k(x_i,x_j)$ points 间通信/耦合强度.

GPR: trainingdata-set $A=\{x_1,...,x_m\}$, observed value-set y_A, prior mean $\mu(x)$, prior mean vector $\boldsymbol\mu(\boldsymbol A)$. $y\sim\mathcal N(0,K_{AA}+\sigma_n^2 I)=\mathcal N(0,K_y)$ **Mean**: $\mu_*(x_*)=\mu(x_*)+k(x_*,A)K_y^{-1}(y_A-\mu_A)$ **Cov**: $k_*(x_*,x')=k(x_*,x')-k(x_*,A)K_y^{-1}k(A,x')$ **Predictive**: $y_*\sim\mathcal N(\mu_*,k_*(x,x')+\sigma_n^2)$

GPRposterior: $y\sim\mathcal N(0,K_{XX}+\sigma_n^2 I)=\mathcal N(0,K_y)$ **Mean**: $\mu'(x)=k(x,X)K_y^{-1}y$ **Cov**: $k'(x,x')=k(x,x')-k(x,X)K_y^{-1}k(X,x')$ **Predictive**: $y_*\sim\mathcal N(\mu',k'+\sigma_n^2)$

Kernels: **Valid Kernel**: 1. PSD: K所有$\lambda_i\ge0$; 2. Closure: k_1+k_2, ck $(c>0)$, $k_1\cdot k_2$, $\exp(k)$ valid; 3. Gram matrix: $x^\top Kx\ge0\ \forall x$ **Linear**: $k(x,x')=x^\top x'+\sigma_0^2$有 rank=1 协方差 matrix(→BLR!). **RBF**: $k=\exp\left(\frac{-\|x-x'\|^2}{2\ell^2}\right)$ smooth σ_f^2: 纵向振幅; ℓ: 横向平滑 **Laplace**: $k=\exp(-\frac r\ell)$ Rough, sharp peaks, C^0 cont **Cosine**: $k=\cos(2\pi\frac rp)$ (Periodic, no decay) **Exponential**: $k=\exp(-\|x-x'\|/\ell)$ rough **Matérn**: $\nu=0.5$ →Exp, $\nu\to\infty$→RBF, ν控制 smoothness **Periodic**: $k=\sigma^2\exp\left(-\frac{2}{\ell^2}\sin^2\left(\frac{\pi|x-x'|}{p}\right)\right)$ **Closure**: k_1+k_2, $k_1\cdot k_2$, $c\cdot k$, $\exp(k)$仍 valid kernel **Stationary**: $k(x,x')=k(x-x')$; **Isotropic**: $k=k(\|x-x'\|)$

Marginal 似然: $\log p(y|X,\theta)=-\frac12 y^\top K_y^{-1}y-\frac12\log|K_y|+C$ Data fit (第一项) vs Complexity (第二项) 多峰非凸: 多个 local 最优 (不同ℓ可能相似似然)

limit 分析: $\sigma_n^2\to\infty$: posterior →先验 (noise 淹没 data) $\ell^2\to\infty$ (Linear): 信号>>noise→最小二乘回归 只有 1train 点: 精确插值该点, elsewhere 高不确定

Sparse GP: Ndata, Minducing: $O(NM^2+M^3)$ vs 标准$O(N^3)$ SoR/FITC/VFE/RFF: 低秩核近似

Precision Matrix: $\Lambda:=\Sigma^{-1}$ (Covariance 的逆) **diag** 项 Λ_{ii}: 条件方差 $\frac{1}{\mathbb V[X_i|X_{-i}]}$ **non-diag** Λ_{ij}: 条件相关性 (给定其他变量); Λ稀疏→条件独立结构 (Graphical Lasso); Gaussian Product: $\Sigma^{-1}=\Sigma_1^{-1}+\Sigma_2^{-1}$ (precision 相加!)

4. VI & ELBO

动机: 积分 $\int p(y_*|w)p(w|D)dw$难算 **Laplace**: 峰值 $\mathcal N(w_{\text{MAP}},H^{-1})$ **VI**: 用 $q(\theta)$近似 $p(\theta|D)$, min KL(q‖p) **MCMC**: 直接采样$w_s\sim p(w|D)$

ELBO: $\mathcal L=\mathbb E_q[\log p(y|\theta)]-\text{KL}(q(\theta)\|p(\theta))$; $\log p(y)=\mathcal L+\text{KL}(q\|p(\theta|y))$ Max ELBO ⟺ Min KL to posterior
等价 **OR**: $\arg\max_q\mathcal L\equiv\arg\min_q\text{KL}(q\|p(\theta|y))\equiv\arg\min_q\mathbb E_q[\log q(\theta)-\log p(y,\theta)]$

Min KL ≡ Max Likelihood: 当 $q(\theta)=\delta(\theta-\hat\theta)$ (point estimate): $\min_\theta\text{KL}(\delta\|p(\theta|y))\equiv\max_\theta p(y|\theta)p(\theta)$ (MAP!) 当 prior uniform: MAP→MLE 非参数族: VI 受q族限制, 无法精确 recover 真实 posterior 例: 真实 posterior multi-modal, q强制 unimodal Gaussian→mode collapse

Gaussian KL: 1维: $\text{KL}(\mathcal N_p\|\mathcal N_q)=\frac12\left[\frac{(\mu_p-\mu_q)^2}{\sigma_q^2}+\frac{\sigma_p^2}{\sigma_q^2}-1-\log\left(\frac{\sigma_p^2}{\sigma_q^2}\right)\right]$ 多维: $\text{KL}=\frac12\left[\text{tr}(\Sigma_q^{-1}\Sigma_p)+(\mu_p-\mu_q)^\top\Sigma_q^{-1}(\mu_p-\mu_q)-d+\log(|\Sigma_q|/|\Sigma_p|)\right]$

Reparam Trick: 公式: $\theta=g(\varepsilon;\lambda)$, $\varepsilon\sim p(\varepsilon)$ (indep λ) $\mathbb E_{q_\lambda}[f(\theta)]=\mathbb E_p[f(g(\varepsilon;\lambda))]$ $\nabla_\lambda\mathbb E_q[f]=\mathbb E_p[\nabla_\lambda f(g(\varepsilon;\lambda))]$ 移入 Grad; **Gaussian**: $\theta=\mu+\sigma\odot\varepsilon$, $\varepsilon\sim\mathcal N(0,I)$ $\nabla_\mu\mathbb E[f(\theta)]=\mathbb E[\nabla_\mu f(\mu+\sigma\varepsilon)]$ $\nabla_\sigma\mathbb E[f(\theta)]=\mathbb E[\nabla_\sigma f(\mu+\sigma\varepsilon)]$ 性质: 连续可微变量; Unbiased; Low variance
Score Fnc: $\mathbb E[f\nabla\log q]$ 离散/连续均可; High variance

Laplace 近似: $q(\theta)=\mathcal N(\hat\theta,\Lambda^{-1})$ $\hat\theta=$MAP; $\Lambda=-\nabla^2\log p(\hat\theta|D)$ (Hessian) Unimodal 分布准确; Multimodal 失效; Mode 处好, elsewhere 过confident

5. BNN & Uncertainty

model: **Prior**: $\theta\sim\mathcal N(0,\sigma_p^2 I)$ **Homoscedastic**: $y|x,\theta\sim\mathcal N(f(x;\theta),\sigma_n^2)$ (固定 noise) **Heteroscedastic**: $y\sim\mathcal N(f_\mu,\exp(f_\sigma))$ (输入依赖) $\sigma^2(x)=\exp(f_\sigma(x;\theta))$ 保证>0

Posterior Log-Density: $\log p(\theta|D)\propto\log p(\theta)+\sum_i\log p(y_i|x_i,\theta)$ $=-\frac{1}{2\sigma_p^2}\|\theta\|^2-\sum_i\left[\frac12\log\sigma^2(x_i)+\frac12\frac{(y_i-\mu(x_i))^2}{\sigma^2(x_i)}\right]$ model 可"blame"noise 但付$\log\sigma$代价

Comparison σ^2: σ_p^2: Prior variance (weight prior) σ_n^2: Aleatoric noise (观测 noise, 固定) $\sigma^2(x)$: Heteroscedastic noise (输入依赖) σ_{epi}^2: Epistemic (model 不确定, data 增加→减少) σ_{ale}^2: Aleatoric (datanoise, data 增加不变); epistemic=认知=可学习; aleatoric=偶然=不可减

Uncertainty: $\mathbb V_{\text{total}}[y]=\underbrace{\mathbb E_\theta[\mathbb V[y|\theta]]}_{\text{aleatoric}}+\underbrace{\mathbb V_\theta[\mathbb E[y|\theta]]}_{\text{epistemic}}$
BLR 闭式: $\sigma_{\text{epi}}^2=x^\top\Sigma_{\text{post}}x$; $\sigma_{\text{ale}}^2=\sigma_n^2$ **GPR 闭式**: $\sigma_{\text{epi}}^2=k'(x,x)$; $\sigma_{\text{ale}}^2=\sigma_n^2$ **MC 近似** (m采样): $\theta_j\sim q(\theta)$ Aleatoric: $\frac1m\sum_j\sigma^2(x,\theta_j)$ Epistemic: $\frac1m\sum_j(\mu(x,\theta_j)-\bar\mu)^2$

MC Dropout: $q_j=p\delta_0+(1-p)\delta_\lambda$ (Bernoulli mask) **Training**: Dropout 开 **Inference**: Dropout 保持开→多次 forward→uncertainty estimate 本质: Variational inference! q是 Bernoulli×Gaussian 混合 近似$p(\theta|D)$但限制在特定族 **vs Gaussian Dropout**: 加性 noise$w+\varepsilon$ vs 乘性 mask $w\cdot m$

方法对比: **SWAG**: SGD 轨迹 avg, $O(d^2)$ 存Σ **Ensembles**: 多 model 独立 train **Calibration**: ECE=$\sum|\text{acc}-\text{conf}|$; Temp Scaling: $\frac zT$

6. Active Learning & BO

Info Gain 目标: $I(S)=I(f_S;y_S)=H[f_S]-H[f_S|y_S]$ **Submodular**: $F(A\cup\{x\})-F(A)\ge F(B\cup\{x\})-F(B)$, $A\subseteq B$ Greedy: $(1-\frac1e)$-approx; NP-hard 最优

对比: **Uncertainty**: $x=\arg\max H[y_x|D]$ Homo 时 OK; Hetero 失效 (混淆 aleatoric/epistemic) **BALD**: $x=\arg\max I(\theta;y_x|D)=H[y_x|D]-\mathbb E_\theta[H[y_x|\theta]]$ 找 model disagreement **Hetero 修正**: $I(f;y|x)=\frac12\log(1+\sigma_{\text{epi}}^2/\sigma_{\text{ale}}^2)$ 考虑 SNR 而非纯 variance

BO Acquisition: **UCB**: $\mu+\beta\sigma$; $\beta=0$→exploit; $\beta\to\infty$→explore **PI**: $\Phi\left(\frac{\mu-f^+}{\sigma}\right)$ 保守 **EI**: $(\mu-f^+)\Phi(Z)+\sigma\varphi(Z)$ 平衡 **Thompson**: 采样 $\tilde f\sim p(f|D)$, arg max $\tilde f$

Regret & Info Gain: $R_T=\sum(f^*-f(x_t))$; Sublinear: $\frac{R_T}{T}\to0$ $R_T=O(\sqrt{T\gamma_T})$ for UCB Linear: $\gamma_T=O(d\log T)$ RBF: $\gamma_T=O((\log T)^{d+1})$

7. MDP Foundations

MDP 定义: $(\mathcal S,\mathcal A,P,R,\gamma)$: states, actions, transitions, reward, discount **Policy** $\pi:\mathcal S\to\mathcal A$ (或$\pi(a|s)$ stochastic) **Stationary**: π与时间t无关 **Deterministic**: $\pi(s)$ 单值; **Stochastic**: $\pi(a|s)$ prob 分布

Value Fnc: $V^\pi(s)=\mathbb E^\pi\left[\sum_{t=0}^\infty\gamma^tR_t|s_0=s\right]$ $Q^\pi(s,a)=\mathbb E^\pi\left[\sum_{t=0}^\infty\gamma^tR_t|s_0=s,a_0=a\right]$ **V&Q**: $V^\pi(s)=\sum_a\pi(a|s)Q^\pi(s,a)$ $Q^\pi(s,a)=R(s,a)+\gamma\sum_{s'}P(s'|s,a)V^\pi(s')$

Bellman Expectation Eq: $V^\pi(s)=\sum_a\pi(a|s)\left[R(s,a)+\gamma\sum_{s'}P(s'|s,a)V^\pi(s')\right]$ **Q value 版**: $Q^\pi(s,a)=R(s,a)+\gamma\sum_{s'}P(s'|s,a)\sum_{a'}\pi(a'|s')Q^\pi(s',a')$ **Matrix**: $\boldsymbol v^\pi=(\boldsymbol I-\gamma\boldsymbol P^\pi)^{-1}\boldsymbol r^\pi$ ($O(n^3)$求解)

BOE: **Bellman** 算子: γ-contraction in $\|\cdot\|_\infty$ $V^*(s)=\max_a\left[R(s,a)+\gamma\sum_{s'}P(s'|s,a)V^*(s')\right]$; $Q^*(s,a)=R(s,a)+\gamma\sum_{s'}P(s'|s,a)\max_{a'}Q^*(s',a')$; **V&Q**: $V^*(s)=\max_aQ^*(s,a)$ $\pi^*(s)=\arg\max_aQ^*(s,a)$

Optimal Policy 定理: π^* optimal $\iff$ greedy w.r.t. own V^π $\pi^*(s) = \arg\max_a \sum_{s'} P(s'|s,a) V^*(s')$ 有限 MDP+$\gamma < 1$: 存在 deterministic stationary π^*

PI vs VI 对比: **Policy Iter**: (1) Eval: 解 LSE $V^\pi = (I - \gamma P^\pi)^{-1} r^\pi$ 精确 (2) Improve: $\pi' = \arg\max_a Q^\pi$ greedy Fewer iters; $O(n^3)$/iter; 收敛到 **exact** π^* 单调性: $V^{\pi_{k+1}} \geq V^{\pi_k}$ 严格改进 **Value Iter**: $V_{k+1}(s) = \max_a \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a) V_{k(s')}\right]$ More iters; $O(n^2 m)$/iter; 收敛到 ε-optimal 收敛: $\|V_{k+1} - V_k\|_\infty < \varepsilon \to V_k \approx V^*$

Reward 变: **Scaling**: $R' = \alpha R$ ($\alpha > 0$) $\to \pi^*$不变, $V' = \alpha V$ 平移: $R' = R + c \to \pi^*$可能变! $V' = V + \frac{c}{1-\gamma}$ $c > 0$+$\gamma \to 1$→偏好长轨迹 **Potential-based**: $F = \gamma\varphi(s') - \varphi(s)$, $R' = R + F \to \pi^*$不变

POMDP: 概念: POMDP 不可直接用 VI/PI, 需转 Belief-MDP (连续状态) **Belief**: $b_t(x) = \mathbb{P}(X_t = x | y_{1:t}, a_{1:t-1})$ **Bayes Filter**: $b_{t+1} \propto o(y_{t+1}|x) \sum_{x'} P(x|x', a_t) b_t(x')$ Belief-state MDP: $\rho(b,a) = \mathbb{E}_{x \sim b}[r(x,a)]$

8. Tabular RL

Q-Learning: Off-policy, Model-free;

$$Q_{\text{new}} = (1-\alpha) Q_{\text{old}} + \alpha\left[r + \gamma \max_{a'} Q_{\text{old}(s',a')}\right]$$

notation: Q_t=处理t个样本后; $Q_0 = 0$ (或 optimistic init) 用 **max**: 理想最优a' (off-policy!) **Convergence**: Robbins-Monro ($\sum \alpha_t = \infty$, $\sum \alpha_t^2 < \infty$) + 所有(s,a)访问无限次 $\boldsymbol{Q}_{t+1}(s,a) = \boldsymbol{Q}_t(s,a) + \alpha[r + \gamma \max_{a'} \boldsymbol{Q}_t(s',a') - \boldsymbol{Q}_t(s,a)]$若$(s,a)$ visited;otherwise $Q_t(s,a)$不动.

SARSA (On-policy, Model-free): $Q_{t+1}(s,a) = Q_t(s,a) + \alpha[r + \gamma Q_t(s',a') - Q_t(s,a)]$ 用实际$\boldsymbol{a}'$: policy 执行的 action (on-policy!) 更保守; a'来自ε-greedy/π

TD Learning (Policy Eval): $V_{t+1}(s) = V_t(s) + \alpha[r + \gamma V_t(s') - V_t(s)]$ **As SGD**: $\ell = \frac{1}{2}[V(s) - (r + \gamma V(s'))]^2$ $\nabla_V \ell = (V(s) - r - \gamma V(s')) \cdot 1$

Model-based (学 MDP): $\hat{P}(s'|s,a) = \frac{N(s'|s,a)}{N(s,a)}$ (count visits) $\hat{R}(s,a) = \frac{\sum_{\text{visits}} r}{N(s,a)}$ 然后用$\hat{P}, \hat{R}$做 VI/PI **vs Model-free**: Q-learning 直接学Q, 不估P, R

Exploration: ε**-greedy**: prob ε random **Optimistic Init**: $Q_0(s,a) = \frac{R_{\max}}{1-\gamma}$ (乐观探索) **Rmax**: unknown $(s,a) \to R_{\max}$; PAC 保证 **H-UCRL**: 乐观选最强 model (OFU 原则)

9. Deep RL

DQN: Off-policy. **Target Net** θ^-: 每C步更新→稳定 **Exp Replay**: buffer 随机采样→打破时间相关 $\mathcal{L} = (r + \gamma \max_{a'} Q_{\theta^-}(s',a') - Q_\theta(s,a))^2$ **Double DQN**: $a^* = \arg\max Q_\theta$; 用$Q_{\theta^-}(s',a^*)$评估 减 maximization bias (Q-learning 过估计)

Policy∇Thrm:

$$\nabla_\theta J = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) G_t\right]$$

, deduction: $\nabla \log P(\tau|\theta) = \sum_t \nabla \log \pi(a_t|s_t)$ $P(\tau) = \mu(s_0) \prod \pi(a_t|s_t) \prod P(s_{t+1}|s_t, a_t)$ dynamics$P(s'|s,a)$抵消 (对θ求导为 0) $G_t = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'}$ (reward-to-go, 因果律!)

REINFORCE (On-policy): $\theta \leftarrow \theta + \alpha \sum_t \nabla \log \pi_\theta(a_t|s_t) G_t$ **MC** 估计: 完整轨迹τ计算 G_t **High variance**(引入 baseline)

Baseline (Variance Reduction): $\nabla J = \mathbb{E}[\sum \nabla \log \pi (G_t - b(s_t))]$ Unbiased if b与a_t无关! **Proof**: $\mathbb{E}_a[b(s) \nabla \log \pi(a|s)] = b(s) \nabla \sum_a \pi(a|s) = 0$ **Optimal**: $b(s) = V^\pi(s)$ (若∇近似常数) **Advantage**: $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$ $\nabla J = \mathbb{E}[\sum \nabla \log \pi \cdot A]$

Actor-Critic: **Actor**: $\pi_\theta(a|s)$; **Critic**: $V_\varphi(s)$ 或 $Q_\varphi(s,a)$ $\nabla J \approx \mathbb{E}[\nabla \log \pi(a|s)(Q(s,a) - V(s))]$ Critic bootstrap→减 variance 但引入 bias (若V_φ不准) **A2C**: Advantage Actor-Critic, on-policy

DDPG(连续 $\mathcal{A}$): **Deterministic policy**: $a = \mu_\theta(s)$ (非 prob!) **train** 加噪: $a_{\text{explore}} = \mu_\theta(s) + \mathcal{N}$ (OUnoise 或 Gaussian) 无 noise→无探索 (deterministic policy 缺陷) **Actor**: $\nabla_\theta J = \mathbb{E}[\nabla_a Q_\varphi(s,a)|_{a=\mu} \nabla_\theta \mu_\theta(s)]$ **Critic**: $\left(r + \gamma Q_{\varphi^-}(s', \mu_{\theta^-}) - Q_\varphi\right)^2$ **Off-policy**: Exp Replay+Target Nets **vs MPC**: MPC 硬算H步$G = \sum r$; DDPG 用 TD $G = r + \gamma Q$

∇ Estimator Bias-Var:

Method	Bias	Var	适用
MC (G_t)	Unbiased	High	完整轨迹
TD bootstrap	Biased	Low	单步
Baseline $G - b$	Unbiased	Lower	b与a无关
Actor-Critic	Biased	Low	Critic 不准时
Reparam	Unbiased	Low	连续可微
Score/REINFORCE	Unbiased	High	通用

Critic bias: 若$V_\varphi \approx V^\pi$准确则 unbiased **Baseline** 条件: $b(s)$只依赖 state, 不依赖 action!

Advanced: **PPO**: Clip$\left(\frac{\pi}{\pi_{\text{old}}}\right)$ 限制 update 幅度; On-policy **SAC**: Entropy regularization $+\lambda H(\pi)$; Off-policy **TRPO**: $\max \mathbb{E}\left[\left(\frac{\pi}{\pi_{\text{old}}}\right) A\right]$ s.t. KL $\leq \delta$

RL Algo Check: **On-policy**: SARSA, REINFORCE, A2C, PPO (data 来自当前π) **Off-policy**: Q-learn, DQN, DDPG, SAC, TD3 (用旧 data/buffer) **Model-based**: Rmax, H-UCRL, PETS, Dyna-Q (学 P, R) **Model-free**: Q-learn, PG, DQN (直接学Q/π) **OFU** (乐观探索): Rmax, H-UCRL (未知→高奖励) **Gradient** 估计: Score Fnc (REINFORCE): High var, 离散/连续均可;Reparam (DDPG): Low var, 需连续可微 **MCTS**: 蒙特卡洛树搜索, 模拟轨迹→UCB 选择 **MPC**: Model Predictive Control, horizon H→执行第 1 步→replan **MCTS vs MPC**: 都需 model; MPC 确定性规划, MCTS 随机搜索

10. Diffusion Models

Def: **Forward**: data→noise (固定q, 无学习) **Backward**: noise→data (学 p_λ) 隐变量 model: $x_{1:T}$ latents, x_0 data

Forward: $q(x_t|x_{t-1}) = \mathcal{N}\left(\sqrt{1-\beta_t} x_{t-1}, \beta_t I\right)$ $x_t = \sqrt{1-\beta_t} x_{t-1} + \sqrt{\beta_t} \varepsilon_t$ Define: $\alpha_t = 1 - \beta_t$; $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ **Closed-Form Marginal**: $q(x_t|x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t} x_0, (1-\bar{\alpha}_t) I)$ $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t} \varepsilon$ $t \to T$: $\bar{\alpha}_T \to 0$, $x_T \sim \mathcal{N}(0, I)$

Backward: $p_\lambda(x_{t-1}|x_t) = \mathcal{N}(\mu_\lambda(x_t, t), \sigma_t^2 I)$ Prior: $p(x_T) = \mathcal{N}(0, I)$ **generate**: $x_T \sim \mathcal{N}(0, I)$ → 迭代 denoise → x_0

Forward Posterior: $q(x_{t-1}|x_t, x_0) = \mathcal{N}\left(\tilde{\mu}_t, \tilde{\beta}_t I\right)$ (给定 x_0, x_t可算!) $\tilde{\mu}_t = \frac{\sqrt{\bar{\alpha}_{t-1}} \beta_t x_0 + \sqrt{\alpha_t}(1-\bar{\alpha}_{t-1}) x_t}{1-\bar{\alpha}_t}$ $\tilde{\beta}_t = \frac{(1-\bar{\alpha}_{t-1})\beta_t}{1-\bar{\alpha}_t}$

ELBO & noise 预测: $\mathcal{L} = \text{const} - \sum_{t=2}^{T} \text{KL}(q(x_{t-1}|x_t, x_0) \| p_\lambda(x_{t-1}|x_t))$ 两 Gauss 间 KL $\propto \|\mu_1 - \mu_2\|^2$ **Issue**: 直接预测μ_λ不稳定(目标依赖x_0) **Solution**: 预测 **noise** ε! 从$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t} \varepsilon$反解: $\tilde{\mu}_t = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \varepsilon\right)$ **simple** $\mathcal{L}$: $L_{\text{simple}} = \mathbb{E}_{t, x_0, \varepsilon}[\|\varepsilon - \varepsilon_\lambda(x_t, t)\|^2]$ **train**=predict noise; Backward=denoise;

train & 采样: **Train**: 采样(x_0, t, ε) → 算x_t → $\nabla \|\varepsilon - \varepsilon_\lambda\|^2$ **Sample**: $x_T \sim \mathcal{N}(0, I)$ For $t = T, \ldots, 1$: $z \sim \mathcal{N}(0, I)$ if $t > 1$ else $z = 0$ $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \varepsilon_\lambda(x_t, t)\right) + \sigma_t z$

与 **BNN** 连接: **Latent var model**: $x_{1:T}$ 类似 BNN hidden layers, **Reparam** 应用: $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t} \varepsilon$, where εreparametrization 变量; **VI** 框架: Max ELBO → Min KL($q \| p_\lambda$) Forward q已知 → Backward p_λ待学

变体: **LDM**: VAE, latent space diffusion (Stable Diffusion) **DDIM**: 确定性($z = 0$), 加速 20-50 步 **Cond**: $\varepsilon_\lambda(x_t, t, c)$; Classifier-Free Guidance

速查 Quick Ref **Hoeffding**:$\mathbb{P}\left(|\hat{X} - \mathbb{E}[X]| \geq \varepsilon\right) \leq 2\exp\left(-2n\frac{\varepsilon^2}{(b-a)^2}\right)$ 仅适用 **iid** data; MCMC 有自相关不适用 **Bellman**: Expectation 用π求和; Optimality 用 max; $V = \sum \pi Q$; $Q = R + \gamma PV$ **PI vs VI**: PI 精确 Eval+少 iters; VI 单步+多 iters; 都$O(n^3)$或$O(n^2 m)$ **Q-learn vs SARSA**: Off 用max vs On 用实际a'; 前者过估计后者保守 **PG** 关键: 动态抵消→$\nabla = \sum \nabla \log \pi G_t$; Baseline 减方差不 bias (与$a$无关!) **DDPG**: 确定$\mu$→必须加噪探索(OU/Gaussian); Off-policy+连续动作 **Diffusion**: Forward 固定加噪→Backward 学 denoise; 预测ε而非μ; ELBO=VI 框架 **Uncertainty**: Epi=model(data↑→↓); Ale=noise(data↑不变); Total=Ale+Epi **Info**: Never Hurts $H[X|Y] \leq H[X]$; $I(X; Y, Z) = I(X; Y) + I(X; Z|Y)$ **Kernel** 验证: PSD (特征值≥ 0); Closure (和/积/指数); Gram $x^\top K x \geq 0$ **MAP vs MLE**: MAP=MLE+正则; Gauss prior→L2; Laplace→L1 **ELBO**: Max $\mathcal{L} \iff$ Min KL$(q\|p)$ $\iff$ Max 似然 (point estimate 时) **Reparam**: $\theta = \mu + \sigma\varepsilon$ → Grad 移入$\mathbb{E}$ → Low var; Score 高 var **VI** 局限: 受q族限制→无法精确 recover (如 multimodal→unimodal) **RL** 分类: On=SARSA/REINFORCE/A2C/PPO; Off=Q/DQN/DDPG/SAC; Model-based=Rmax/PETS **OFU**: 乐观探索=未知高奖励; Rmax/H-UCRL; Optimistic Init $Q_0 = \frac{R_{\max}}{1-\gamma}$ **MCTS vs MPC**: 树搜索 vs 确定规划; 都需 model; MPC 执行 1 步 replan

Linear kernel GP ≡ BLR; Uniform prior 时: MAP=MLE; VI 能精确 recover 任意 posterior ✗ (受q族限制); Entropy 正则化→偏好 stochastic/uniform; 边缘似然关于 hyperparams 是凸的 ✗ (通常多峰非凸!); 预测 noise ≡ 预测 mean(等价但 noise 更 stable), Forward process 无需 learning. Contextual Bandit = MDP with $|S| = 1$, $\gamma = 0$. $\frac{\partial}{\partial\theta}\log|K| = \text{tr}(K^{-1}\partial\frac{K}{\partial\theta})$ $\frac{\partial}{\partial\theta}(y^\top K^{-1} y) = -y^\top K^{-1} \frac{\partial K}{\partial\theta} K^{-1} y$

Dict **A2C**:Advantage Actor-Critic; **BALD**:Bayesian Active Learning by Disagreement; **BLR**:Bayesian Linear Reg; **BNN**:Bayesian NN; **BO**:Bayesian Opt; **DDIM**:Denoising Diffusion Implicit; **DDPG**:Deep Deterministic PG; **DDPM**:Denoising Diffusion Prob; **DQN**:Deep Q-Net; **ECE**:Expected Calibration Error; **EI**:Expected Improvement; **ELBO**:Evidence Lower Bound; **GPR**:GP Regression; **H-UCRL**:Hoeffding-UCB RL; **LDM**:Latent Diffusion; **LOTV**:Law of Total Var; **MAP**:Max A Posteriori; **MCTS**:Monte Carlo Tree Search; **MI**:Mutual Info; **MLE**:Max Likelihood Est; **MPC**:Model Predictive Control; **NLL**:Negative Log-Likelihood; **OU**:Ornstein-Uhlenbeck; **PETS**:Probabilistic Ensemble w/ Trajectory Sampling; **PI**:Prob of Improvement; **PPO**:Proximal Policy Opt; **RBF**:Radial Basis Fnc; **RFF**:Random Fourier Features; **SAC**:Soft Actor-Critic; **SNR**:Signal-Noise Ratio; **SWAG**:Stoch Weight Avg Gaussian; **TD**:Temporal Diff; **UCB**:Upper Confidence Bound; **VI**:Variational Inference; ln2≈0.693; ln3≈1.099; 1/e≈0.368; √2≈1.414; (1-1/e)≈0.632; $aX + bY \sim \mathcal{N}(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2)$ (indep); Bern $[p]$, $[p(1-p)]$; Binomial(n, p $[np]$, $[np(1-p)]$; Poisson(λ) $[\lambda]$, $[\lambda]$; Exp(λ $[\lambda e^{-\lambda x}]$, $[\frac{1}{\lambda}]$, $[\frac{1}{\lambda^2}]$;

RL Algo Property:

Algo	On/Off	Model	Data Eff	Complexity	Bias/Var
Q-learn	Off	Free	High	$O(\|S\|\|A\|)$	Unbiased
SARSA	On	Free	Low	$O(\|S\|\|A\|)$	Unbiased
DQN	Off	Free	High	Func Approx	Biased(FA)
DDPG	Off	Free	High	Func Approx	Biased(FA)
REINFORCE	On	Free	Low	Func Approx	Unbiased/HiVar
A2C	On	Free	Med	Func Approx	Biased(Critic)
PPO	On	Free	Med	Func Approx	Biased(Clip)
SAC	Off	Free	High	Func Approx	Biased(Entropy)
Rmax	Both	Based	Low	$O(\|S\|^3)$	Unbiased
PETS	Off	Based	High	Ensemble	Biased(Model)

∇ Estimator:Score Fnc (REINFORCE): Unbiased, High Var, 离散/连续; Reparam (DDPG/SAC): Biased(若 FA), Low Var, 连续 only; Baseline: Unbiased iff 与 action 无关