Dict **BALD**:Bayesian Active Learning by Disagreement; **BLR**:Bayesian Linear Reg; **BN**:Bayesian Network; **BNN**:Bayesian NN; **BO**:Bayesian Opt; **BP**:Belief Propagation; **CPD**:Cond Prob Dist; **DAG**:Directed Acyclic Graph; **DBE**:Detailed Balance Eq; **DDIM**:Denoising Diffusion Implicit Models; **DDPG**:Deep Deterministic PG; **DDPM**:Denoising Diffusion Prob Models; **DQN**:Deep Q-Net; **ECE**:Expected Calibration Error; **EI**:Expected Improvement; **ELBO**:Evidence Lower Bound; **GP**:Gaussian Process; GPR: GP Regression; **HMM**:Hidden Markov Model; **KF**:Kalman Filter; **KL**:Kullback-Leibler; **LDM**:Latent Diffusion; **LOTV**:Law of Total Var; **MALA**:Metropolis-Adjusted Langevin; **MAP**:Max A Posteriori; **MCMC**:Markov Chain MC; **MDP**:Markov Decision Process; **MH**:Metropolis-Hastings; **MI**:Mutual Info; **MLE**:Max Likelihood Est; **MPE**:Most Probable Explanation; **PF**:Particle Filter; **PI**:Prob of Improvement; **POMDP**:Partially Observable MDP; **RBF**:Radial Basis Fnc; **RFF**:Random Fourier Features; **SGLD**:Stoch Grad Langevin Dyn; **SWAG**:Stoch Weight Avg Gaussian; **TD**:Temporal Diff; **UCB**:Upper Confidence Bound; **VE**:Var Elimination; **VI**:Variational Inference;

Probability Fundamentals

Axioms: $\mathbb{P}(\Omega)=1$; $\mathbb{P}(A)\geq 0$; Disjoint: $\mathbb{P}\left(\bigcup_i A_i\right)=\sum_i \mathbb{P}(A_i)$ **Product**: $\mathbb{P}(X_{1:n})=\mathbb{P}(X_1)\prod_{i=2}^n \mathbb{P}(X_i|X_{1:i-1})$ **Sum**: $\mathbb{P}(X)=\sum_y \mathbb{P}(X,y)$ **Bayes**: $\mathbb{P}(X|Y)=\frac{\mathbb{P}(Y|X)\mathbb{P}(X)}{\mathbb{P}(Y)}$ **Cond Indep**: $X\perp Y|Z \leftrightarrow \mathbb{P}(X,Y|Z)=\mathbb{P}(X|Z)\mathbb{P}(Y|Z)$

Gaussian: $\mathcal{N}(x;\mu,\Sigma)=\frac{\exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right)}{\sqrt{(2\pi)^d|\Sigma|}}$ **Marginal**: $X_A\sim\mathcal{N}(\mu_A,\Sigma_{AA})$ **Conditional**: $X_A|X_B\sim\mathcal{N}\left(\mu_{A|B},\Sigma_{A|B}\right)$ $\mu_{A|B}=\mu_A+\Sigma_{AB}\Sigma_{BB}^{-1}(x_B-\mu_B)$ $\Sigma_{A|B}=\Sigma_{AA}-\Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}$ **Linear**: $Y=MX\sim\mathcal{N}(M\mu,M\Sigma M^\top)$ **Sum**: indep $X+X'\sim\mathcal{N}(\mu+\mu',\Sigma+\Sigma')$

$\mathbb{E}$, Var, Cov, Info: $\mathbb{E}[AX+b]=A\mathbb{E}[X]+b$; **Tower**: $\mathbb{E}_Y[\mathbb{E}_X[X|Y]]=\mathbb{E}[X]$ $\mathrm{Var}[X]=\mathbb{E}[(X-\mathbb{E}[X])^2]$; $\mathrm{Cov}[X,Y]=\mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]$ $\mathrm{Var}[X+Y]=\mathrm{Var}[X]+\mathrm{Var}[Y]+2\mathrm{Cov}[X,Y]$ **LOTV**: $\mathrm{Var}[X]=\mathbb{E}[\mathrm{Var}[X|Y]]+\mathrm{Var}[\mathbb{E}[X|Y]]$ **Entropy**: $H[p]=-\mathbb{E}_p[\log p(x)]$; Gauss $H=\frac{1}{2}\log\left((2\pi e)^d\det\Sigma\right)$ **KL**: $\mathrm{KL}(p\|q)=\mathbb{E}_p\left[\log\frac{p}{q}\right]\geq 0$; need $\mathrm{supp}(q)\subseteq\mathrm{supp}(p)$ **Forward** $\mathrm{KL}(p\|q)$: mean-seeking 覆盖; **Reverse** $\mathrm{KL}(q\|p)$: mode-seeking 过 confident **MI**: $I(X;Y)=H[X]-H[X|Y]=H[Y]-H[Y|X]\geq 0$, symmetric **Cond MI**: $I(X;Y|Z)=H[X|Z]-H[X|Y,Z]$ **Gauss MI**: $I[X;Y]=\frac{1}{2}\log\det(I+\sigma_n^{-1}\Sigma)$ for $Y=X+\varepsilon$ Gaussian prior→L2; Laplace prior→L1 **MLE**: $\hat\theta_{\mathrm{MLE}}=\arg\max_\theta\sum_i\log p(y_i|x_i,\theta)$ **MAP**: $\hat\theta_{\mathrm{MAP}}=\arg\min_\theta\underbrace{-\log p(\theta)}_{\text{reg}}+\underbrace{\ell_{\mathrm{nll}}}_{\text{fit}}$

BLR:= GP with Linear 核 $k(x,x')=x^\top x'$

Model: $y=w^\top x+\varepsilon$, $\varepsilon\sim\mathcal{N}(0,\sigma_n^2)$; **Prior**: $w\sim\mathcal{N}(0,\sigma_p^2 I)$, L_2正则/weight decay; **Posterior**: $w|X,y\sim\mathcal{N}(\mu,\Sigma)$ where $\Sigma^{-1}=\sigma_n^{-2}X^\top X+\sigma_p^{-2}I$; $\mu=\sigma_n^{-2}\Sigma X^\top y$, Σ只依赖X, 不依赖y

Prediction: $y^*|x^*,X,y\sim\mathcal{N}\left(x^{*\top}\mu,x^{*\top}\Sigma x^*+\sigma_n^2\right)$ $\mu \Longleftrightarrow$`RidgeReg`解(=MAP解), Σ则对应其Hessian的逆. `MAP=Ridge with` $\lambda=\frac{\sigma_n^2}{\sigma_p^2}$; Online update: $O(nd^2)$

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 $A=\{x_1,\ldots,x_m\}$, 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: set $A=\{x_1,\ldots,x_m\}$, $y\sim\mathcal{N}(0,K_{AA}+\sigma_n^2 I)=\mathcal{N}(0,K_y)$ **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**: $y^*\sim\mathcal{N}(\mu^*,k^*(x,x')+\sigma_n^2)$

Kernels: **Linear**: $k(x,x')=x^\top x'+\sigma_0^2$ **RBF**: $k=\exp\left(\frac{-\|x-x'\|^2}{2\ell^2}\right)$ smooth 无限可微 **Laplace**: $k=\exp\left(-\frac{r}{\ell}\right)$ Rough, sharp peaks, C^0 cont **Cosine**: $k=\cos\left(2\pi\frac{r}{p}\right)$ (Periodic, no decay) **Exponential**: $k=\exp(-\|x-x'\|/\ell)$ rough **Matérn**: $\nu=0.5$→Exp, $\nu\to\infty$→RBF, ν控制 smoothness **HyperParam** :ℓ length-scale, x-axis wiggle speed; σ_f^2 (amplitude, y-axis scale). **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 Lik: $\log p(y|X)=-\frac{1}{2}y^\top K_y^{-1}y-\frac{1}{2}\log\det(K_y)+C$ Balance: Data fit(前) vs Complexity(后)

Approx $O(n^3)$ → lower: **RFF**: $k(x-x')\approx\varphi(x)^\top\varphi(x')$, $O(nm^2+m^3)$ Bochner: stationary kernel ↔ Fourier of non-neg measure **Inducing Pts**: subset $m\ll n$ points for approx, $O(NM^2)$(LoRA)

Variational Inference, ELBO

Motiv: 考虑$p(y^*|x^*,D)=\int p(y^*|w)p(w|D)dw$ 不同 approx 处理 intractable 积分方式不同: Laplace(峰值) $p(w|D)\approx\mathcal{N}(w_{\mathrm{MAP}},-H^{-1})$1次 Hessian 且可微; VI `min` ELBO, 用$q(\cdot)$近似Z; MC 直接 sample $w_s\approx p(w|D)$ unbiasd 地估 近似$p(\theta|D)$ with $q(\theta|\lambda)$ by min $\mathrm{KL}(q\|p)$, $q(\cdot)$是自定义的 approx 分布. 注意识别$\log(p(y))$对于$\mathbb{E}_{q(\theta)}$无关, $\mathbb{E}_q[\log p(y)]=\log p(y)$为 const. Recall Bayese $p(Z)p(X|Z)=p(X)p(Z|X)$化简 ELBO 时.

ELBO: $\mathcal{L}=\mathbb{E}_q[\log p(y|\theta)]-\mathrm{KL}(q(\theta)\|p(\theta))$ 且 $\underbrace{\log p(y)}_{\text{Evidence}}=\mathcal{L}+\underbrace{\mathrm{KL}(q\|p(\theta|y))}_{\text{Error }\geq 0}$. Max ELBO ⟺ Min KL to posterior **Derivation**: Jensen: $\log\mathbb{E}_q\left[\frac{p}{q}\right]\geq\mathbb{E}_q\left[\log\frac{p}{q}\right]\Rightarrow\arg\max_q\mathcal{L}\equiv\arg\min_q\mathrm{KL}(q\|p(\theta|y))\equiv\arg\min_q\mathbb{E}_q[\log q(\theta)-\log p(y,\theta)]$

KL of Gaussian $\mathrm{KL}(\mathcal{N}_p\|\mathcal{N}_q)=\frac{1}{2}\left[\mathrm{tr}(\Sigma_q^{-1}\Sigma_p)+(\mu_p-\mu_q)^\top\Sigma_q^{-1}(\mu_p-\mu_q)-d+\log\frac{\det[\Sigma_q]}{\det[\Sigma_p]}\right]$ Σ_q^{-1}: precision 矩阵, trace 项算匹配程度(linear), log det 算熵差异(log). $\mathrm{KL}(\mathcal{N}_p\|\mathcal{N}_q)=\frac{1}{2}\left[\frac{(\mu_p-\mu_q)^2}{\sigma_q^2}+\frac{\sigma_p^2}{\sigma_q^2}-1-\log\frac{\sigma_p^2}{\sigma_q^2}\right]$ 1dim 时. **Product**: $\mathrm{KL}(Q_XQ_Y\|P_XP_Y)=\mathrm{KL}(Q_X\|P_X)+\mathrm{KL}(Q_Y\|P_Y)$

Reparam Trick: **Formula**: $\theta=g(\varepsilon;\lambda)$, $\varepsilon\sim p(\varepsilon)$ (Indep of λ) $\mathbb{E}_{q_\lambda}[f(\theta)]=\mathbb{E}_p[f(g(\varepsilon;\lambda))]$ $\nabla_\lambda\mathcal{L}\approx\frac{1}{L}\sum\nabla_\lambda f(g(\varepsilon;\lambda))$ e.g. **Gaussian**: $z=\mu+\sigma\odot\varepsilon$, $\varepsilon\sim\mathcal{N}(0,I)$ **property**: 需 Continuous(differentiable)变量(Auto-diff), **unbiased**, low-variance;对比 `REINFORCE`之 $\mathbb{E}[f(\theta\nabla_\lambda\log q(\theta))]$, 虽然都是∇的 **unbiased** estimate, 但后者连续/离散 variable 均适用, 高方差用 baseline 来降.

Laplace Approx: motiv: 用二次 fit $\log p(x)$, 适用 unimodal(not multimodal). $q(\theta)=\mathcal{N}\left(\hat\theta,\Lambda^{-1}\right)$ $\hat\theta=$ MAP; $\Lambda=-\nabla^2\log p\left(\hat\theta|D\right)$ (Hessian) Good at mode, overconfident elsewhere 考虑 Gaussian 就是 unimode, 二次型 log-density, LaplaceApprox 能精确 recover 之 $\log P(Z|X)=\nabla P(X,Z)-\nabla\log Z$而 $\nabla\log Z=0$, 找 mode 无需Z.

Bayesian Neural Networks

Model: Prior: $\theta\sim\mathcal{N}(0,\sigma_p^2 I)$ **Homoscedastic**: $y|x,\theta\sim\mathcal{N}(f(x;\theta),\sigma^2)$ fixed noise **Heteroscedastic**: $y\sim\mathcal{N}(f_\mu(x;\theta),\exp\{f_\sigma(x;\theta)\})$ input-dependent noise

Hetero NLL: $-\log p(y|x,\theta)=C+\frac{1}{2}\left[\log\sigma^2(x)+\frac{[(y-\mu(x)]^2)}{\sigma^2(x)}\right]$ Model can "blame" noise but pays $\log\sigma$ penalty 防 collapse

MAP for BNN: $\hat\theta_{\mathrm{MAP}}=\arg\min\frac{1}{2\sigma_p^2}\|\theta\|^2+\frac{1}{2\sigma_n^2}\sum_i[y_i-f(x_i;\theta)]^2$ Weight decay = Gaussian prior

Prediction: $p(y^*|x^*,D)\approx\frac{1}{m}\sum_{j=1}^m p\left(y^*|x^*,\theta^{(j)}\right)$, $\theta^{((j)\}}\sim q$ MC approx of posterior predictive

Aleatoric vs Epistemic: $\sigma_{\text{total}}=\sigma_e+\sigma_a\Leftrightarrow\mathrm{Var}(y)=\mathrm{Var}(\mathbb{E}[y|\theta])+\mathbb{E}[\mathrm{Var}(y|\theta)]$ "Var of Means"+"Mean of Vars"也解析解: 如 **BLR** 中 param 后验, $\sigma_{\text{epi}}^2=x^{*\top}\Sigma_{\text{post}}x^*$ 如 **GPR** 中 param 后验, $\sigma_{\text{epi}}^2=k_*(x^*,x^*)$; $\sigma_{\text{ale}}^2=\sigma_n^2$ (constant noise) 近似解:(MC/Ensemble, m此 sampling), θ_j: 某近似 posterior. Aleatoric(data noise): $\frac{1}{m}\sum_j\sigma^2\left(x^*,\theta^{(j)}\right)$ Epistemic(model uncertainty): $\frac{1}{m}\sum_j\left[\mu\left(x^*,\theta^{(j)}\right)-\bar\mu\right]^2$ where $\bar\mu=\frac{1}{m}\sum_j\mu\left(x^*,\theta^{(j)}\right)$

MC Dropout: $q_j(\theta_j)=p\delta_0(\theta_j)+(1-p)\delta_{\lambda_j}(\theta_j)$ Test 时 keep dropout⟶多次 forward passes⟶uncertainty estimates

SWAG: Store running avg of SGD iterates: μ,Σ Space: $O(d^2)$ covariance vs $O(Td)$ all models

Calibration: Goal: Confidence ≈ Accuracy **ECE**: $\sum\frac{|B_m|}{n}|\mathrm{acc}(B_m)-\mathrm{conf}(B_m)|$ **Temp Scaling**: $\frac{z}{T}$ on logits; $T>1$⟶less confident

Active Learning

Objective: $I(S)=I(f_S;y_S)=H[f_S]-H[f_S|y_S]$ NP-hard; Greedy gives $(1-\frac{1}{e})$-approx (submodular, monotone)

Strategies: **Uncertainty Sampling**: $x=\arg\max H[y_x|D]$ Cannot distinguish aleatoric vs epistemic **BALD**: $x=\arg\max I(\theta;y_x|D)=H[y_x|D]-\mathbb{E}_\theta[H[y_x|\theta]]$ Finds where models **disagree** about y_x **Hetero**: $x=\arg\max\{\sigma_{\text{epistemic}}^2/\sigma_{\text{aleatoric}}^2\}$

Submodular: $F(A\cup\{x\})-F(A)\geq F(B\cup\{x\})-F(B)$ for $A\subseteq B$ Diminishing returns; MI is submodular

Bayesian Optimization

Regret: $R_T=\sum_{t=1}^T(f^*-f(x_t))$ Goal: sublinear $R_T/T\to 0$

Acquisition Fncs: **UCB**: $x_{t+1}=\arg\max\left[\mu_{t(x)}+\beta_t\sigma_{t(x)}\right]$ $\beta_t=0$: pure exploit; $\beta_t\to\infty$: uncertainty sampling Regret: $R_T=O\left(\sqrt{T\gamma_T}\right)$ **PI**: $\mathrm{PI}(x)=\Phi\left(\frac{\mu(x)-f(x^+)}{\sigma(x)}\right)$ **EI**: $\mathrm{EI}(x)=(\mu-f^+)\Phi(Z)+\sigma\varphi(Z)$, $Z=\frac{\mu-f^+}{\sigma}$ **Thompson**: Sample $\tilde f\sim p(f|D_t)$, pick $\arg\max\tilde f$

Info Gain γ_T: Linear: $\gamma_T=O(d\log T)$ RBF: $\gamma_T=O\left((\log T)^{d+1}\right)$ Matérn($\nu>\frac{1}{2}$): $\gamma_T=O\left(T^{\frac{d}{2\nu+d}}(\log T)^{2\frac{\nu}{2\nu+d}}\right)$

MDP & Bellman & Hoeffding 不等式 & Concentration

MDP: $(\mathcal{S},\mathcal{A},P,R,\gamma)$: states, actions, $P_{sa}(s')$, reward $R(s,a)$ or $R(s)$, discount $\gamma\in[0,1)$. Policy $\pi:\mathcal{S}\to\mathcal{A}$. **Value fnc**: $V^\pi(s)=\mathbb{E}\left[\sum_{t\geq 0}\gamma^t R(s_t)\mid s_0=s,\pi\right]=R(s)+\gamma\sum_{s'\in S}P_{s,\pi(s)}(s')V^\pi(s')$ (follow π) **Q-fnc**: $Q^\pi(s,a)=R(s,a)+\gamma\sum_{s'}P_{sa}(s')V^\pi(s')=\sum_{s'}P_{sa}(s')\left[R(s,a,s')+\gamma\sum_{a'}\pi(a'|s')Q^\pi(s',a')\right]$

BOE: $V^*(s)=R(s)+\max_a\gamma\sum_{s'}P_{sa}(s')V^*(s')$ $Q^*(s,a)=R(s,a)+\gamma\sum_{s'}P_{sa}(s')\max_{a'}Q^*(s',a')$ **Bellman** 期望 **Eq**: $V^\pi(s_t)=\mathbb{E}_\pi[r_t+\gamma V^\pi(s_{t+1})]=\sum_a\pi(a|s)\sum_{s'}P_{sa}(s')\left[R(s,a,s')+\gamma V^{\pi(s')}\right]$ Q-fnc 版: $V^{\pi(s)}=\sum_a\pi(a|s)Q^{\pi(s,a)}$, 则: $Q^\pi(s,a)=\mathbb{E}\left[R_t+\gamma\sum_{a'}\pi(a'\mid s_{t+1})Q^\pi(s_{t+1},a')\right]$ **Optimal** policy: $\pi^*(s)=\arg\max_a\sum_{s'}P_{sa}(s')V^*(s')$ Policy opt ⇔ greedy w.r.t. V^π. Bellman op γ-contraction in ℓ_∞. **Matrix**: $\boldsymbol{v}^\pi=(\boldsymbol{I}-\gamma\boldsymbol{T}^\pi)^{-1}\boldsymbol{r}^\pi$ $(O(n^3))$

Horizon: **Infinite** ($\gamma<1$): $V=\sum_{t=0}^\infty\gamma^t r_t$ converges. **Finite** (H steps): $V=\sum_{t=0}^{H-1}\gamma^t r_t$. Even $\gamma=1$ converges(finite sum).Conversion: $H\approx\frac{1}{1-\gamma}$.

Horizon Bound: $X_i\in[a,b]$ i.i.d: $\mathbb{P}\left(|\hat X-\mathbb{E}[X]|\geq\varepsilon\right)\leq 2\exp\left(-2n\frac{\varepsilon^2}{(b-a)^2}\right)$;say$R\in[0,1]$: CI $(1-\delta)\to|\hat\mu-\mu|\leq\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}}$;**UCB bonus**: 不确定性$\sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n}}$; set $\delta_t=\frac{1}{t^2}\to\sqrt{\frac{2\ln t}{n}}$; Hoeffding bound 只适用 **iid** data, MCMC samples 有自相关性 thus 不适用;

MDP & RL Foundations

Bellman Eqs: **Expectation**: $V^{\pi(s)}=R(s,\pi(s))+\gamma\sum_{s'}P(s'|s,\pi(s))V^{\pi(s')}$ **Optimality**: $V^*(s)=\max_a\left[R(s,a)+\gamma\sum_{s'}P(s'|s,a)V^*(s')\right]$ $Q^*(s,a)=R(s,a)+\gamma\sum_{s'}P(s'|s,a)\max_{a'}Q^*(s',a')$ **Matrix**: $\boldsymbol{v}^\pi=(\boldsymbol{I}-\gamma\boldsymbol{P}^\pi)^{-1}\boldsymbol{r}^\pi$

Bellman's Thrm: π^* optimal iff greedy w.r.t. own V^π: $\pi^*(s)=\arg\max_a Q^*(s,a)$

PI & VI: **Policy Iter**: (1)Eval V^π exactly(solve LSE), (2) $\pi\to$greedy. Fewer iters, $O(n^3)$/iter. **Value Iter**: $V\to\max_a[r+\gamma PV]$. More iters, $O(n^2m)$/iter. Both converge to optimal; VI gives ε-optimal

POMDP: Belief-state MDP: reward $\rho(b,a)=\mathbb{E}_{x\sim b}[r(x,a)]$ **Belief**: $b_t(x)=P(X_t=x|y_{1:t},a_{1:t-1})$ **Bayes Filter**: $b_{t+1}(x)\propto o(y_{t+1}|x)\sum_{x'}P(x|x',a_t)b_t(x')$

Tabular RL

Model-based: $\hat P(x'|x,a)=\frac{N(x'|x,a)}{N(a|x)}$, $\hat r(x,a)$ =avg rewards Converges but needs many samples

Q-Learning **(Off-policy)**: $Q(s,a)\leftarrow Q(s,a)+\alpha(r+\gamma\max_{a'}Q(s',a')-Q(s,a))$ Uses max (ideal best a'); off-policy, model-free

SARSA (On-policy): $Q(s,a) \leftarrow Q(s,a) + \alpha(r + \gamma Q(s',a') - Q(s,a))$ Uses actual a' from policy; on-policy

TD Learning: $V(s) \leftarrow V(s) + \alpha(r + \gamma V(s') - V(s))$ **As SGD**: $\ell = \frac{1}{2}[V(s) - (r + \gamma V(s'))]^2$ Converges if Robbins-Monro: $\sum \alpha_t = \infty, \sum \alpha_t^2 < \infty$

Exploration: **ε-greedy**: prob ε random, else best **Optimistic Init**: $Q = \frac{R_{\max}}{1-\gamma}$ **Rmax**: unknown$(s,a) \to R_{\max}$, PAC guarantee

Deep RL

DQN: $\mathcal{L} = (r + \gamma \max_{a'} Q_{\theta^-}(s',a') - Q_\theta(s,a))^2$ **Target Net** θ^-: stabilize; **Experience Replay**: break correlation **Double DQN**: selection θ, eval θ^-; reduces overestimation

Policy Gradient: $\nabla_\theta J = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_t \nabla \log \pi_\theta(a_t|s_t) G_t\right]$ $\nabla \log P(\tau) = \sum_t \nabla \log \pi(a_t|s_t)$ (dynamics cancel!) **REINFORCE**: MC estimate, high variance **Baseline**: $G_t - b(s_t)$, $b = V(s)$ optimal; unbiased

Actor-Critic: **Actor**: $\pi_\theta(a|s)$; **Critic**: $V^\varphi(s)$ or $Q^\varphi(s,a)$ $\nabla J \approx \mathbb{E}[\nabla \log \pi(a|s)(Q(s,a) - V(s))]$ Critic bootstrap 减 variance 但引入 bias

Advanced: **TRPO**: $\max \mathbb{E}\left[\left(\frac{\pi_\theta}{\pi_{\text{old}}}\right) A^{\pi_{\text{old}}}\right]$ s.t. KL $\le \delta$ **DDPG**: continuous actions, deterministic $\mu_\theta(s)$ **Adv Fnc**: $A^{\pi(s,a)} = Q^{\pi(s,a)} - V^{\pi(s)}$

Performance Diff Lemma(Episodic): $V^{\pi'}(s_0) - V^\pi(s_0) = \sum_{t=0}^{H-1} \mathbb{E}_{s\sim d_t^{\pi'}, a\sim\pi'}[Q_t^\pi(s,a) - V_t^\pi(s)] = \sum_{t=0}^{H-1} \mathbb{E}_{s\sim d_t^{\pi'}, a\sim\pi'}[A_t^\pi(s,a)]$

Policy ∇ Theory, PG Thrm & Estimators

Trajectory: $\nabla_\theta \log P(\tau|\theta) = \sum_{t=1}^H \nabla_\theta \log \pi_\theta(a_t|s_t)$ Gradient Environment dynamics $P(s'|s,a)$ and $\mu(s_1)$ cancel out! **Trajectory**: $\tau = (s_1, a_1, s_2, a_2, \ldots, s_H, a_H)$, $P(\tau) = \mu(s_0)\prod_{t=0}^{T-1} \pi(a_t|s_t) P(s_{t+1}|s_t,a_t)$ **Log**: $\log P(\tau|\theta) = \log \mu(s_1) + \sum_{t=1}^H \log \pi_\theta(a_t|s_t) + \sum_{t=1}^H \log P(s_{t+1}|s_t,a_t)$

PG Thrm: $\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^T \nabla_\theta \log \pi_\theta(a_t|s_t) \cdot G_t\right]$ where $G_{t:H} = \sum_{t'=t}^H \gamma^{t'-t} r_{t'}$ (return/reward-to-go),or 写成 $R(\tau)$. ∇增加 a_t的 prob, 降 others. 还可 rewrite 原始 PG Thrm as, recall $Q^{\pi(s_t,a_t)} = \mathbb{E}_{\tau\sim\pi}[G_{t:H}|s_t,a_t]$, $\mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^\infty \gamma^t \nabla_\theta \log \pi_\theta(a_t|s_t)(Q_\theta(s_t,a_t) - \mathbf{0})\right]$ **REINFORCE**: MC estimate of G_t, high variance. 采样 m轨迹 τ_i from π_{θ_k}, 算 unbiased ∇ 估计后 update $\pi_{\theta_k} += \alpha_k \widehat{\nabla_\theta J}$ $\widehat{\nabla_\theta J} = \frac{1}{m}\sum_{i=1}^m \left(\sum_{t=1}^H \gamma^t R_{i,t}\right)\left(\sum_{t=1}^H \nabla_\theta \log \pi_{\theta_k}\right)(a_{i,t}|s_{i,t})$ 良τ_i充当 MLE in SupvL, 需多 samples, 没用 Mrkv 性. **Actor-Critic**: Replace G_t with Q^π or A^π estimated by critic. Actor: via $\nabla_\theta \log \pi_\theta A_t$ update π_θ; Critic: learn $\hat{V}_\omega(s_t)$or $\hat{A}_\omega(s_t,a_t)$ to Approx baseline. $\nabla_\theta J(\theta) \approx \sum_t \nabla_\theta \log \pi_\theta(\cdot|\cdot)\left(r_t + \gamma \hat{V}_\omega(s_{t+1}) - \hat{V}_\omega(s_t)\right)$, biased 若 $\hat{V}_\varphi^\pi$不准. $\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_t \nabla_\theta \log \pi_\theta(\cdot|\cdot)[r(s_{i,t},a_{i,t}) + \gamma \hat{V}_\varphi^\pi(s_{i,t+1} - \hat{V}_\varphi^\pi(s_{i,t})]$. 低方差, critic 近似 long-term return.

Softmax ∇: $\pi_\theta(a|s) = \left(e^{\beta Q_\theta(s,a)}\right)/\left(\sum_b e^{\beta Q_\theta(s,b)}\right)$ policy 表达式, Softmax $\nabla \log \pi$: $\nabla_\theta \log \pi_\theta(a|s) = \beta\left(\nabla_\theta Q_\theta(s,a) - \sum_b \pi_\theta(b|s) \nabla_\theta Q_\theta(s,b)\right) = \beta(\nabla_\theta Q_\theta(s,a) - \mathbb{E}_{b\sim\pi}[\nabla_\theta Q_\theta(s,b)])$ 后者 as baseline

Baseline Unbiasedness: For any $b(s)$ depending only on state (not action). ∴ State-dependent baseline *never* introduces bias: $\mathbb{E}_{a\sim\pi(\cdot|s)}[b(s)\nabla_\theta \log \pi_\theta(a|s)] = b(s)\nabla_\theta \sum_a \pi_\theta(a|s) = b(s)\cdot 0 = 0$ **With baseline**: $\nabla_\theta J = \mathbb{E}\left[\sum_t \nabla \log \pi(G_t - b(s_t))\right]$ Optimal baseline: $b^*(s) = V^\pi(s)$ (minimizes σ^2, if ∇roughly const)

Diffusion Models

Setup: **Forward**: data→noise (fixed, no learning) **Backward**: noise→data (learned generation) Latent var model: $x_{1:T}$ are latents, x_0 is data

Forward Process: $q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right)$ $x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\varepsilon_t$ Schedule: $\beta_t \in (0,1)$单调增, $\beta_1 \approx 10^{-4}$, $\beta_T \approx 0.02$

Closed-Form Marginal ⭐: Define: $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ $q(x_t|x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I)$ **Reparam**: $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon$, $\varepsilon \sim \mathcal{N}(0,I)$ As $t \to T$: $\bar{\alpha}_T \to 0$, $x_T \sim \mathcal{N}(0,I)$ indep of x_0

Reverse Process: $p_{\lambda(x_{t-1}|x_t)} = \mathcal{N}\left(\mu_{\lambda(x_t,t)}, \sigma_t^2 I\right)$ Prior: $p(x_T) = \mathcal{N}(0,I)$ Generate: sample x_T, iteratively denoise to x_0

Forward Posterior: $q(x_{t-1}|x_t,x_0) = \mathcal{N}\left(\tilde{\mu}_t, \tilde{\beta}_t I\right)$ $\tilde{\mu}_t = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t$ $\tilde{\beta}_t = \frac{(1-\bar{\alpha}_{t-1})\beta_t}{1-\bar{\alpha}_t}$ Key: given x_0, x_t, forward posterior is Gaussian (tractable)

ELBO & Loss: $\mathcal{L} = \text{const} - \sum_{t=2}^T \underbrace{\text{KL}\left(q(x_{t-1}|x_t,x_0) \| p_{\lambda(x_{t-1}|x_t)}\right)}_{L_t}$ Two Gaussians same var: KL $\propto \|\mu_1 - \mu_2\|^2$

⭐**Noise Prediction** : Predict ε instead of μ (more stable): From $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon$: $\tilde{\mu}_t = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon\right)$ **Simple Loss**: $L_{\text{simple}} = \mathbb{E}_{t,x_0,\varepsilon}\left[\|\varepsilon - \varepsilon_{\lambda(x_t,t)}\|^2\right]$

Training Algo: Repeat: sample $x_0 \sim p_{\text{data}}$, $t \sim \text{Unif}\{1,\ldots,T\}$, $\varepsilon \sim \mathcal{N}(0,I)$ $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon$ $\nabla_\lambda \|\varepsilon - \varepsilon_{\lambda(x_t,t)}\|^2$

Sampling Algo: $x_T \sim \mathcal{N}(0,I)$ For $t = T, \ldots, 1$: $z \sim \mathcal{N}(0,I)$ if $t > 1$ else $z = 0$ $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_{\lambda(x_t,t)}\right) + \sigma_t z$

Connection: $\varepsilon_{\lambda(x_t,t)} \approx -\sqrt{1-\bar{\alpha}_t}\nabla_{x_t} \log q(x_t)$ **Denoising = Score matching**

Variants: **LDM**: diffusion in VAE latent space, more efficient **DDIM**: deterministic sampling, fewer steps **Cond Gen**: $\varepsilon_{\lambda(x_t,t,c)}$, Classifier-Free Guidance: $\tilde{\varepsilon} = (1+w)\varepsilon_{\lambda(x_t,t,c)} - w\varepsilon_{\lambda(x_t,t)}$

QuickCheck:

- **VI**: Approx posterior via ELBO. Laplace MAP, Reparam for grad.
- **MCMC**: Sample posterior. MH accept/reject, Gibbs coordinate, Langevin uses ∇.
- **GP**: Prior over fncs, closed-form posterior. RBF smooth, Matérn tunable.
- **BNN**: Prior on weights, MC predictive. Aleatoric=data noise, Epistemic=model.
- **Active**: Max MI, BALD for disagreement, submodular→greedy$(1-\frac{1}{e})$.
- **BO**: UCB balance explore/exploit, EI expected gain, Thompson sample.
- **BN**: DAG factorization, d-sep for indep, BP exact on trees.
- **KF**: Linear Gaussian, Kalman gain trades predict vs observe.
- **Diffusion**: Forward=noise, Backward=denoise, train predict ε.
- **On/Off**: On=SARSA,REINFORCE,PPO; Off=Q-learn,DQN,SAC
- **Bellman**: $V = R + \gamma PV$;