1 Probability

Bayes Rule: $P(X|Y) = \frac{P(X,Y)}{P(Y)} = \frac{P(X)P(Y|X)}{\sum_{X=x} P(X=x)P(Y|X=x)}$

1.1 Gaussian

$\mathcal{N}(x;\Sigma,\mu) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$, $O(n^2)$ var

$\mathbf{Cov(x_i,\ x_j)}$: $\mathbb{E}[(x_i-\mu_i)(x_j-\mu_j)]$

Marginal: $X_A = [X_{i_1},...,X_{i_k}] \sim \mathcal{N}(\mu_A, \Sigma_{AA})$

Conditional: $p(X_A|X_B = x_B) = \mathcal{N}(\mu_{A|B}, \Sigma_{A|B})$

$\mu_{A|B} = \mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \mu_B)$, $\Sigma_{A|B} = \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}$

Times: $M \in \mathcal{R}^{m\times d}, MX \sim \mathcal{N}(M\mu_X, M\Sigma_{XX}M^T)$

Add: $X_1 + X_2 \sim \mathcal{N}(\mu_1+\mu_2, \Sigma_1+\Sigma_2), X_1 + s \sim \mathcal{N}(\mu_1+s, \Sigma_1)$

2 Matrix manipulation

$\nabla(\mathbf{a^T w}) = a$ $\nabla(\mathbf{w^T B w}) = 2Bw$

$\sum(y_i - w^Tx_i)^2 = (y-Xw)^T(y-Xw) = ||y-Xw||^2$

Woodbury: $U(VU+I) = (UV+I)U$

$(A+xx^T)^{-1} = A^{-1}\frac{(A^{-1}x)(A^{-1}x)^T}{1+x^TA^{-1}x}$

3 Bayesian Linear Regression

Ridge: $y \approx w^Tx$, $\hat{w} = \arg\min_w \sum_i^n (y_i - w^Tx_i)^2 + \lambda||w||_2^2$

Analytical: $\hat{w} = (X^TX+\lambda I)^{-1}X^Ty$

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

Lhood: $P(y|w,x) = \prod_i^n P(y_i|w,x_i) = \prod_i^n \mathcal{N}(y_i; w^Tx_i, \sigma_n^2)$

$w = \arg\max_w P(w|X,y) = \arg\max_w (\frac{1}{Z})P(w|X)\prod_i P(y_i|w,x_i)$

$P(\hat{w}|X,y) = \mathcal{N}(\hat{w}; (X^TX + \frac{\sigma_n^2}{\sigma_p^2}I)^{-1}X^Ty, (\sigma_n^{-2}X^TX + \sigma_p^{-2}I)^{-1})$

Predict: $P(y^*|X,y,x^*) = \int p(y^*|x^*,w)p(w|x_{1:n},y_{1:n})dw$

$= \mathcal{N}(\hat{\mu_w}^T x^*, x^{*T}\hat{\Sigma_w} x^* + \sigma_n^2)$, in $O(nd^2)$

Epistemic: Lack of data, the part $x^{*T}\hat{\Sigma}x^*$

Aleatoric: Irreducible noise from observation, the part σ_n^2

Hyperparas: $\lambda = \frac{\sigma_n^2}{\sigma_p^2}$ via CV, $\sigma_n^2 \approx \frac{1}{n}\sum_i(y_i - w^Tx_i)^2$

4 Gaussian Process

Introduce non-linearity: $f(x) = w^T\phi(x) \Rightarrow$ dim. explosion

Kernel: $x_i^T x_j \Rightarrow \phi(x)^T\phi(x) = k(x_i, x_j)$

F-space: $f = [f_{i:n}] = Xw$, $w \sim \mathcal{N}(0, \sigma_p^2 I)$, $f \sim \mathcal{N}(0, \sigma_p^2 XX^T)$

Predict: $y^* = f^* + \epsilon \sim \mathcal{N}(0, \sigma_p^2\hat{X}\hat{X}^T + \sigma_n^2 I)$, $\hat{X}^T = [X^T, x^*]$

GP: $f \sim GP(\mu(.), k(.,.))$, $\mu: X \to \mathbb{R}, k: X\times X \to \mathbb{R}$

$P(f|X_A, y_A) = GP(f; \mu(x) + k_{x,A}(K_{AA}+\sigma_n^2 I)^{-1}(y-\mu_A),$

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

Sample: $f \sim \mathcal{N}(0,K), K = LL^T, \epsilon \sim \mathcal{N}(0,I) \Rightarrow f = L\epsilon$

Model select: $\hat{\theta} = \arg\max_\theta \int p(y_{train}|X_{train}, f, \theta)p(f|\theta)df$

$= \arg\max \int \mathcal{N}(y; f(x), \sigma_n^2)\mathcal{N}(f; 0, K_f(\theta)) = \arg\max$

$\mathcal{N}(y; 0, K_f(\theta) + \sigma_n^2 I) = \arg\min -\frac{1}{2}y^TK_y(\theta)^{-1}y - \frac{1}{2}\log|K_y(\theta)|$

Performance: Parallel, Local, Kernel approx, $O(n^3)$

Fourier: Shift-invariant kernel features BL, $O(nm^2 + m^3)$

Bochner: Shift-invariant kernel p.d. $\Leftrightarrow p(\omega)$ non-negative

4.1 Kernel

Correlation: $k(x_1, x_2) = Cov(f(x_1), f(x_2))$

$\forall x, x', k(x,x') = k(x',x)$, p.s.d (positive EV) $x^TK_{AA}x \geq 0$.

Composition: $+, \cdot, k\cdot const., poly(f), \exp(f)$ give again kernels

Stationary: if it holds $k(x,x') = k(x-x')$

Isotropic: $k(x,x') = k(||x-x'||_2)$, implies stationary

Linear: $k(x,x') = x^Tx'$ = Bayesian linear regression

Poly2: $k(x,x') = \phi(x)^T\phi(x'), \phi(x) = [1, x, x^2]$

Exp^2: $k(x,x') = \exp(-||x-x'||_2^2/h^2)$, decay with distance

Exp: $k(x,x') = \exp(-||x-x'||_1/h)$, decay with distance

5 Approximate Inference

Prior: $p(\theta)$, **Likelihood:** $p(y|X) = \prod p(y_i|x_i,\theta)$

Bayesian Posterior: $p(\theta|X,y) = \frac{1}{Z}p(\theta)\prod_{i=1}^n p(y_i|x_i,\theta)$

Prediction: $p(y^*|x^*, x_{1:n}, y_{1:n}) = \int p(y^*|x^*,\theta)p(\theta|x_{1:n}, y_{1:n})$

5.1 Variational Inference

Goal: $p(\theta|y) = \frac{1}{Z}p(\theta,y) \approx q(\theta|\lambda)$

5.2 Laplace Approximation

$\hat{f}(\theta) = \log p(\theta|y) \approx f(\hat{\theta}) + (\theta-\hat{\theta})^Tf(\hat{\theta})' + \frac{1}{2}(\theta-\hat{\theta})^Tf(\hat{\theta})''(\theta-\hat{\theta})$

$q(\theta) = \frac{1}{Z}\exp(\hat{f}(\theta)) \sim \mathcal{N}(\hat{\theta}, \Lambda^{-1}), \Lambda = -\nabla\nabla\log p(\hat{\theta}|y),$

$\hat{\theta} = \arg\max p(\theta|y) \to$ **SGD:** $\theta_{t+1} = \theta_t - \eta_t\frac{1}{m}\sum^m \nabla_\theta l(\theta_t; x_i)$

$p(y^*|x^*, X) = \int p(y^*|x^*,\theta)p(\theta|X,y) = \int p(y^*|f^*)\int p(f^*|\theta)q_\lambda(\theta)$

$\approx \int p(y^*|f^*)\mathcal{N}(f^*; \mu^Tx^*, x^{*T}\Sigma x^*) \leftarrow$ **Predictive distr.**

5.3 KL-Divergence

KL: $KL(q||p) = \int q(\theta)\log\frac{q(\theta)}{p(\theta)}d\theta = \mathbb{E}_{\theta\sim q}[\log\frac{q(\theta)}{p(\theta)}] \geq 0 \leftarrow$ rev.

$KL(p||q) = \frac{1}{2}(tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1\text{-}\mu_0)^T\Sigma_1^{-1}(\mu_1\text{-}\mu_0) - d + ln(\frac{|\Sigma_1|}{|\Sigma_0|}))$

$p \sim \mathcal{N}(\mu_0, I), q \sim \mathcal{N}(\mu_q, I), \to \frac{||\mu_1-\mu_0||^2}{2}$ $|p \sim \mathcal{N}(\mu_{1:d}, [\sigma_{1:d}^2]),$

$q \sim \mathcal{N}(0, \sigma_p^2 I) \to \frac{1}{2}\sum_{i=1}^d(\frac{\sigma_i^2}{\sigma_p^2} + \frac{\mu_i^2}{\sigma_p^2} - 1 - ln(\frac{\sigma_p^2}{\sigma_i^2})$

Entropy: $H(q) = -\int q(\theta)\log q(\theta)d\theta = \mathbb{E}_{\theta\sim q}[-\log q(\theta)]$

Prod: $H(q) = \sum_{i=1}^d H(q_i)$, $H(\mathcal{N}(\mu,\Sigma)) = \frac{1}{2}ln(|2\pi e\Sigma|)$

KL: $q^* \in \arg\min_q KL(q||p) = \arg\min_{q_\lambda}\mathbb{E}_{\theta\sim q}[log\frac{q(\theta)}{\frac{1}{Z}p(\theta,y)}]$

$= \arg\max_q \mathbb{E}_{\theta\sim q}[log(p(\theta,y)] + H(q)$

$= \arg\max_q \mathbb{E}_{\theta\sim q}[log(p(y|\theta)] - KL(q||p(.))$ = ELBO, **BLR** $\downarrow$

$\mathbb{E}_{\theta\sim q_\lambda}[-\sum\log(1+\exp(-y_i\theta^Tx_i))] - \frac{1}{2}\sum^d(\mu_i^2 + \sigma_i^2 - 1 - ln(\sigma_i^2))$

5.4 Repatameterization

$q(\theta|\lambda) = \phi(\epsilon)|\nabla_\epsilon g(\epsilon;\lambda)|^{-1}$, g invertible

$\Rightarrow \nabla_\lambda\mathbb{E}_{\theta\sim q_\lambda}[f(\theta)] = \mathbb{E}_{\epsilon\sim\phi}[\nabla_\lambda f(g(\epsilon;\lambda))]$, unbiased SGD est.

diag-Gauss: $\theta = C\epsilon + \mu, \Sigma = CC^T, \phi(\epsilon) = \mathcal{N}(\epsilon; 0, I)$

ELBO: $\nabla_\lambda\mathbb{E}_{\theta\sim q(.|\lambda)}[log(p(y|\theta)] - \nabla_\lambda KL(q_\lambda||p(.))$

$= \nabla_{C,\mu}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}[log(p(y|C\epsilon+\mu, X)] - \nabla_{C,\mu}KL(q_{C,\mu}||p(.))$

$\approx n\cdot\frac{1}{m}\sum_j^m\nabla_{C,\mu}log(p(y_{ij}|C\epsilon^{(j)}+\mu, x_{ij}) - \nabla_{C,\mu}KL$, unbiased

6 Bayesian Logistic Regression

Prior: $p(\theta) = \mathcal{N}(0, \sigma_p^2\cdot I)$, **Lhood:** $p(y|X,w) = \prod\sigma(y_i\cdot w^Tx_i)$

$p(y|x,w) = Ber(y; \sigma(w^Tx))$

Laplace: $\hat{w} = \arg\max_w p(w|y) = \text{am}_w \log p(w) + \log p(y|w)$

$= \arg\min_w \frac{1}{2\sigma_p^2}||w||_2^2 + \sum_{i=1}^n\log(1+\exp(-y_iw^Tx_i))$ (use SGD)

SGD: $w = w(1-2\lambda\eta_t) + \eta_t yx\frac{1}{1+exp(yw^Tx)}(\leftarrow\hat{P}(Y=-y|w,x))$

$\Lambda = -\nabla\nabla\log p(\hat{w}|X,y) = X^Tdiag([\sigma(\hat{w}^Tx_i)(1-\sigma(\hat{w}^Tx_i))]_i)X$

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

7 Monte Carlo Sampling

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

$\mathbb{E}_{\theta\sim p(.|X,y)}[p(y^*|x^*,\theta)] \approx \frac{1}{m}\Sigma_i^m p(y^*|x^*,\theta^{(i)}), \theta^{(i)} \sim p(\theta|X,y)$

LargeNums: $\mathbb{E}_P[f(X)] \approx \frac{1}{N}\sum_i^N f(x_i)$ **Hoeffding:** f in $[0,C]$,

$P(|\mathbb{E}_P[f(X)] - \frac{1}{N}\sum_i^N f(x_i)| > \epsilon) \leq 2\exp(-2N\epsilon^2/C^2)$

Approx: $\frac{1}{Z}Q(x) = P(x)$, MC seq. $X_{1:N}$ with stationary $P(x)$

Prior $P(X_1)$, transitions $P(X_{t+1}|X_t)$ i.i.d of t

Ergodic: Reach all states from everywhere in exactly t steps

Statnary: $\lim_\infty P(X_N = x) = \pi(x)$, implied by$\uparrow$, $P(X_1)$ egal

Sample: $x_1 \sim P(X_1), x_n \sim P(X_N|X_{N-1} = x_{N-1})$

Goal: $\pi(x) = \frac{1}{Z}Q(x)$, need to specify $P(x|x')$

Balance: $\frac{1}{Z}Q(x)P(x'|x) = \frac{1}{Z}Q(x')P(x|x') \Rightarrow \pi(x) = \frac{1}{Z}Q(x)$

7.1 Metropolis Hastings

1) Proposal: Given $X_t = x$, sample $x' \sim R(X'|X = x)$

2) Accept: w/ probability $\alpha = \min(1, \frac{Q(x')R(x|x')}{Q(x)R(x'|x)})$

Theorem: Stationary is $Z^{-1}Q(x)$

Gibbs: $t..\infty: i \sim Uniform$, update $x_i = P(X_i|x_{-i})$

$P(X_i = x_i|x_{-i}) = \frac{1}{Z}Q(X_i = x_i, x_{-i}) = \frac{Q(x_{1:n})}{\sum_{x_i}Q(X_i=x_i, x_{-1})}$

$\uparrow$ sat. balance equation, practical variant not. $P(X_i|x_{-i})$ eff.

LLN (Ergodic): $lim_{N\to\infty}\frac{1}{N}\sum_i^N f(x_i) = \sum_{x\in D}\pi(x)f(x)$

Expectations: $\mathbb{E}[f(X)] \approx \frac{1}{T-t_0}\sum_{\tau=t_0+1}^T f(X^{(\tau)})$, t_0 burn-in

General RV: $Q(x) = \frac{1}{Z}\exp(-f(x))), f(x) > 0$

$p(\theta|X,y) = \frac{1}{Z}p(\theta)p(y|X,\theta) = \frac{1}{Z}\exp(\text{-}[-\log p(\theta) - \log p(y|X,\theta)])$

Bay.Logistic Reg.: $f(\theta) = \lambda||\theta||^2 + \sum log(1+exp(-y_i\theta^Tx_i))$

Hastings: $\alpha = \min(1, \frac{R(x|x')}{R(x'|x)}\exp(f(x) - f(x')))$ $|\downarrow$ b.c. sym.

$R(x'|x) = \mathcal{N}(x'; x; \tau I) \Rightarrow \alpha = \min\{1, \exp(f(x) - f(x'))\}$

$R(.|.) = \mathcal{N}(x'; x\text{-}\tau\nabla f(x); 2\tau I)$, eval. $f(x)$ both steps

8 Bayesian Deep Learning

Neural Net: $f(x; w) = \varphi(W_1\varphi(W_2(..\varphi(W_lx)))$

Basic Unit: $v_i^{(l)} = \varphi(w_i^{(l)T} v^{(l-1)})$

tanh(z) $= \frac{\exp(z)-\exp(-z)}{\exp(z)+\exp(-z)}$ **reLu(z)** $= \max(z, 0)$

Bayesian NN: $p(y|x,\theta) = \mathcal{N}(y; f_1(x,\theta), f_2(x,\theta))$

MAP: $\hat{\theta} = \arg\min_\theta \lambda||\theta||^2 + \sum_{i=1}^n \log\sigma^2(x_i,\theta) + \frac{(y_i-\mu(x_i,\theta))^2}{\sigma^2(x_i,\theta)}$

Predict: $p(y^*|x^*, x_{1:n}, y_{1:n}) = \mathbb{E}_{\theta\sim p(.|X,y)}[p(y^*|x^*,\theta)]$
$\approx \frac{1}{m}\sum_j^m \mathcal{N}(y^*; \mu(x^*,\theta^{(j)}), \sigma^2(x^*,\theta^{(j)}))$

Mean: $\mathbb{E}[y^*|X,y,x^*] \approx \bar{\mu}(x^*) = \frac{1}{m}\sum_j^m \mu(x^*,\theta^{(j)})$

$Var[y^*|X,y,x^*] = Var[\mathbb{E}[y^*|X,y,x^*]] + \mathbb{E}[Var[y^*|X,y,x^*]]$
$= \frac{1}{m}\sum_j^m \sigma^2(x^*,\theta^{(j)}) + \frac{1}{m}\sum_j^m(\mu(x^*,\theta^{(j)}) - \bar{\mu}(x^*))^2$

8.1 MCM

Predict: $p(y^*|X,y,x^*) \approx \frac{1}{T}\sum_j^T p(y^*|x^*,\theta^{(j)})$, $\theta^{(i)}$ as NN W

Approx: $q(\theta|\mu_{1:d}, \sigma^2_{1:d}), \mu = \frac{1}{T}\sum_j^T \theta^{(j)}, \sigma^2 = \frac{1}{T}\sum_j^T(\theta^{(j)} - \mu)^2$

softmax(f): $p(y|x,\theta) = p_y = \frac{\exp(f_y)}{\sum_j^c \exp(f_j)}$, $softmax(f+\epsilon)$

9 Active Learning

Note: GP Posterior covariance does not depend on y

Entropy: $H(X) = \mathbb{E}_{x\sim p(x)}[-\log p(x)], H(X|Y)$
$= \mathbb{E}_{y\sim p(y)}[H(X|Y=y)], H(X) + H(Y|X) = H(X,Y)$

Gauss: $X \sim \mathcal{N}(\mu,\Sigma), H(x) = \frac{1}{2}\log(2\pi e)^d|\Sigma|$

Mutual Info: $I(X;Y) = H(X) - H(X|Y), I(X;Y|Z) =$
$I(X|Z) - H(X|Y,Z)$, symmetric

Gauss: $X \sim \mathcal{N}(\mu,\Sigma), Y = X + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma_n^2 \cdot I)$
$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(\epsilon)$
$= \frac{1}{2}\log(2\pi e)^d|\Sigma + \sigma_n^2 I| - \frac{1}{2}\log(2\pi e)^d|\sigma_n^2 I| = \frac{1}{2}\log|I + \sigma_n^{-2}\Sigma|$

Target: $\max F(S) = I(f; y_S) = H(f) - H(f|y_S)$, w/ $S \subseteq D$

Greedy (Homoscedasdic): $x_{t+1} = \arg\max_{x\in D} F(S_t \cup \{x\})$
$= \arg\max_{x\in D} I(f; y_{S_t+x}) - I(f; y_{S_t}) = \arg\max_{x\in D} H(f) - H(f|y_{S_t+x}) - H(f) + H(f|y_{S_t}) =^{const.\sigma_n^2} \arg\max_{x\in D}\sigma_t^2(x)$

Greedy (Heteroscedastic): $x_{t+1} = \arg\max_{x\in D}\frac{\sigma_f^2(x)}{\sigma_n^2(x)}$

Performance: Cons-factor approx. (near opt.), submodular

Classification: $x_{t+1} = \arg\max_{x\in D} H(Y|x, x_{1:n}, y_{1:n})$
$= \arg\max_{x\in D} -\sum_y \log p(y|x, x_{1:n}, y_{1:n})$

10 Bayesian Optimization

Given: Noisy black-box f, choose $x_1, .., x_T| \downarrow$ if $\frac{R_T}{T} \to 0$

Regret: $R_T = \sum_t^T(\max_x f(x) - f(x_t)) \Rightarrow \max f(x_t) \to f(x^*)$

UpConfidence: f at least highest lower bound (f enclosed)

GP-UCB: $x_t = \arg\max_{x\in D}\mu_{t-1}(x) + \beta_t\sigma_{t-1}(x)$

Reget: $\frac{1}{T}\sum^T[f(x^*) - f(x_t)] = O(\sqrt{\frac{\gamma_T}{T}}), \gamma_T = \max_{\le T} I(f; y_s)$

Lin: $\gamma_T = O(d\log T)$, **Exp**$= O((logT)^{d+1})$, **Matrn**$= O(TlogT)$

Thompson: $\hat{f} \sim P(f|x_{1:n}, y_{1:n}), x_{t+1} \in \arg\max\hat{f}(x)$

11 Markov Decision Process

MDP: State X, Action A, Transition $P(x'|x,a)$, Reward $r(x,a)$

Policy: $\pi: X \to A$, $P(X_{t+1} = x'|X_t = x) = P(x'|x,\pi(x))$, **or** $\pi: X \to P(A)$, $P(X_{t+1} = x'|X_t = x) = \sum_a \pi(a|x)p(x'|x,a)$

Expectation: $J(\pi) = \mathbb{E}[\sum_{t=0}^\infty \gamma^t r(X_t, \pi(X_t))]$, for MC: $X_0, ..$

$\mathbf{V^\pi(x)}$$= J(\pi|X_0 = x) = r(x,\pi(x)) + \gamma\sum_{x'} p(x'|x,\pi(x))V^\pi(x')$

Exact solution: $V^\pi = (I - \gamma T^\pi)^{-1} r^\pi$, w/ $\gamma < 1$, P/r known

Fixed Point: for $t = 1:T$ do $V^\pi = r^\pi + \gamma T^\pi V_{t-1}^\pi$
$B^\pi V := r^\pi + \gamma T^\pi V \Rightarrow B^\pi V^\pi = V^\pi$
$||B^\pi V - B^\pi V'|| \le \gamma||V - V'||, ||V_t^\pi - V^\pi|| \le \gamma^t||V_0^\pi - V^\pi||$

Greedy Policy: $\pi^* = \arg\max J(\pi)$, # policies $O(|X|^{|A|})$

w.r.t. V: $\pi_G(x) = \arg\max_a r(x,a) + \gamma\sum_{x'} p(x'|x,a)V(x')$

Bellman: $V^*(x) = \max_a[r(x,a) + \gamma\sum_{x'} p(x'|x,a)V^*(x')]$
$Q^*(x,a) = \mathbb{E}_{x'}[r(x,a) + \gamma\max_{x'} Q^*(x',a')]$

Policy Iteration: 1) $V^\pi(x)$ 2) π_G w.r.t. V^π 3) $\pi = \pi_G$

Convergence: $V^{\pi_{t+1}} \ge V^{\pi_t}$, π^* in $O(n^2 m/(1-\gamma))$

Value iteration: $V_t(x) = \max_a Q_t(x,a)$ until $||V_t - V_{t-1}|| \le \epsilon$
$\forall a, x: Q_t(x,a) = r(x,a) + \gamma\sum_{x'} P(x'|x,a)V_{t-1}(x')$

PI: exact, expensive **vs. VI:** cheap, more iters, $\epsilon-$optimal

11.1 POMDP- Belief state MDP

Observe: $P(Y_{t+1} = y|b_t, a_t) = \sum_{x,x'} b_t(x)P(x'|x,a_t)P(y|x')$
$b_{t+1}(x') = \frac{1}{Z}\sum_x b_t(x)P(X_{t+1} = x'|X_t = x, a_t)P(y_{t+1}|x')$

Reward: $r(b_t, a_t) = \sum_x b_t(x)r(x,a_t)$

12 Reinforcement Learning

Goal: max. discounted rewards $\sum^\infty \gamma^t r(X_t, A_t)$ **Data:** $\tau^{(i)} = (x_0^{(i)}, a_0^{(i)}, r_0^{(i)}, x_1^{(i)}, a_1^{(i)}, r_1^{(i)}, ..), D = \{x_j^{(i)}, a_j^{(i)}, r_j^{(i)}, x_{j+1}^{(i)}\}$

Dilemma: Explore or Exploit?

On-policy: Agent takes actions following a policy π

Off-policy: No specific policy is followed to learn

Model-based: Estimate MDP $(P(x'|x,a), r(x,a))$, optimize policy based on MDP

Model-free: Estimate V directly (Policy gradient, AC)

12.1 Model-based RL

Estimates: Transition probabilities and reward function
$\hat{p}(X_{t+1}|X_t, A) = \frac{Count(X_{t+1}, X_t, A)}{Count(X_t, A)}, \hat{r}(x,a) = \frac{1}{N_{x,a}}\sum R(x,a)$

ϵ_t**-greedy:** Pick random or best action, $\sum\epsilon = \infty, \sum\epsilon^2 < \infty$

$\mathbf{R}_{max}$:

Hoefding: $P(|\mu - \frac{1}{n}\sum_i^n Z_i| > \epsilon) \le 2\exp(-2n\epsilon^2/C^2) = \delta$, $Z \in [0, C]$

$P(|\hat{r}(x,a) - r(x,a)| \le \epsilon) \ge 1 - \delta$, then $n_{x,a} \in O(\frac{R_{max}^2}{\epsilon^2}\log\frac{1}{\delta})$

Performance: w/ prob. $1-\delta$, $\epsilon-$optimal policy in # steps polynomial in $|X|, |A|, T, 1/\epsilon, \log(1/\delta)$ and R_{max}.

Memory: $\hat{p}(x'|x,a)$ $(O(|A||X|^2))$ and $\hat{r}(x,a)$ $(O(|X|^2))$

Computation: Solve MDP every epoch, $O(|X||A|)$ for R_{max}

12.2 Model-free RL

TD: $\hat{V}^\pi(x) = (1-\alpha_t)\hat{V}^\pi(x) + \alpha_t(r + \gamma\hat{V}^\pi(x'))$, w/ (x,a,r,x')

Conditions: $\sum_t \alpha_t = \infty, \sum_t \alpha_t^2 < \infty$, ∞ visits, follow π

$\mathbf{Q^*(x,a)}$$= r(x,a) + \gamma\sum_{x'} p(x'|x,a)V^*(x')$

$\mathbf{V^*(x)}$$= \max_a Q^*(x,a)$, even off-policy

Q-lern: $\hat{Q}^*(x,a) = (1-\alpha_t)\hat{Q}^*(x,a) + \alpha_t(r + \gamma\max_{a'}\hat{Q}^*(x',a'))$

Optimistic: $\hat{Q}^*(x,a) = \frac{R_{max}}{1-\gamma}\prod_t^{T_init}(1-\alpha_t)^{-1}$, perf. as R_{max}

Memory: Store all $Q^*(x,a)$ $(O(|A||X|))$

Cost: (single update) compute $\max_{a'} Q^*(x',a')$, $O(|A|)$

TD-SGD: $l_2(\theta; x, x', r) = \frac{1}{2}(V(x;\theta) - r - \gamma V(x';\theta_{old}))^2$

Bellman Error: $\delta = Q(x,a;\theta) - r - \gamma\max_{a'} Q(x',a';\theta_{old})$

$L(\theta) = \sum_{(x,a,r,x')\in D}(Q(x,a;\theta) - r - \gamma\max_{a'} Q(x',a';\theta^{old}))^2$

$l_2(\theta; x, a, x', r) = \frac{1}{2}(\delta)^2$, $\nabla l_2() = \delta\nabla Q(x,a;\theta)$

Linear: $Q(x,a;\theta) = \theta^T\phi(x,a)$, $\nabla Q(x,a;\theta) = \phi(x,a)$

Con: $a_t = \max_a ...$, can be intractable, algo is slow

Q-iter: Keep $\max_{a'} Q(x',a';\theta^{old})$ for multiple iterations

$L^{DDQN}(\theta) = \sum_{(x,a,r,x')\in D}(r + \gamma Q(x', a^*(\theta); \theta^{old}) - Q(x,a;\theta))^2$,
$a^*(\theta) = \arg\max_{a'} Q(x',a';\theta)$

Policy parameterization: $\pi(x) = \pi(x;\theta)$ $\tau^{(1)}, .., \tau^{(m)} \sim \pi_\theta$
$\theta^* = \arg\max_\theta J(\theta) \approx \arg\max_\theta \frac{1}{m}\sum_m^M\sum_t^T \gamma^t r_t^{(m)}$

Grad: $\nabla J(\theta) = \nabla\mathbb{E}_{\tau\sim\pi_\theta} r(\tau) = \mathbb{E}_{\tau\sim\pi_\theta}[r(\tau)\nabla\log\pi_\theta(\tau)]$
$= \mathbb{E}_{\tau\sim\pi_\theta}[\sum_t^T(r(\tau) - b)\nabla\log\pi_\theta(a_t|x_t;\theta)]$

Unbiased, but very large **variance**

Baseline: $b(\tau_{0:t-1}) = \sum_{t'=0}^{t-1}\gamma^{t'} r_{t'}$, $G_t = \sum_{t'=t}^T \gamma^{t'-t} r'_t$

Reinforce: repeat: generate τ, $\theta = \theta + \eta G_t\nabla_\theta\log\pi(a_t|x_t;\theta)$
$\nabla J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[\sum_t^T G_t\nabla\log\pi_\theta(\tau)]$
$\nabla J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[\sum_t^T(G_t - b_t(x_t))\nabla\log\pi_\theta(\tau)]$
f.e. mean $b_t(x_t) = \frac{1}{T}\sum_0^T G_t$

13 Actor Critic

$A^\pi(x,a) = Q^\pi(x,a) - V^\pi(x) = Q^\pi(x,a) - \mathbb{E}_{a'\sim\pi(x)}[Q^\pi(x,a')]$

Policy Grad: $\nabla J(\theta) = \mathbb{E}_{(x,a)\sim\pi_\theta}[Q(x,a;\theta_Q)\nabla\log\pi(a|x;\theta_\pi)]$

A2C: $\theta_\pi \leftarrow \theta_\pi + \eta[Q(x,a;\theta_Q) - V(x;\theta_V)]\nabla\log\pi(a|x;\theta_\pi)$

Note: On policy, use expectation over policy

13.1 Parameterised Policies

$L(\theta_Q) = \sum_{(x,a,r,x')\in D}(r + \gamma Q(x', \pi(x';\theta_\pi); \theta_Q^{old}) - Q(x,a;\theta_Q))^2$

Target: $\pi_G(x) = \arg\max_a Q^\pi(x,a) = \arg\max_a A^\pi(x,a)$

Objective: $\theta_\pi^* \in \arg\max_\theta \mathbb{E}_{x\sim\mu}[Q(x,\pi(x;\theta_\pi);\theta_Q]$

Grad: $\nabla_\theta J_\mu(\theta) = \mathbb{E}_{x\sim\mu}[\nabla_a Q(x,a)|_{a=\pi(x;\theta_\pi)}\nabla_{\theta_\pi}\pi(x;\theta_\pi)]$

DDPG: act after $a = \pi(x;\theta_\pi) + \epsilon, \epsilon \sim \mathcal{N}(0,\lambda I)$
$y = r + \gamma Q(x', \pi(x', \theta_\pi^{old}), \theta_Q^{old})$, update
$\theta_Q \leftarrow \theta_Q - \eta\nabla\frac{1}{|B|}\Sigma(Q(x,a;\theta_Q) - y)^2$, $\theta_Q^{old} \leftarrow (1-\rho)\theta_Q^{old} + \rho\theta_Q^{old}$
$\theta_\pi \leftarrow \theta_\pi + \eta\nabla\frac{1}{|B|}\Sigma Q(x,\pi(x;\theta_\pi);\theta_Q)$, $\theta_\pi^{old} \leftarrow (1-\rho)\theta_\pi^{old} + \rho\theta_\pi^{old}$