ETHz PAI Cheat Sheet

Probability & BLR

- Gaussian

Definition: Gaussian (Normal) distribution

$$\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\Sigma)=\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\Big(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big)$$

$P(\mathcal{N}|\mathcal{N})\sim\mathcal{N}(\mu_{A|B},\Sigma_{A|B})$

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

$P(M\mathcal{N})\sim\mathcal{N}(\mu_Y,\Sigma_Y)$

$$\mu_Y=M\mu_X$$
$$\Sigma_{YY}=M^\top\Sigma_{XX}M$$

$P(\mathcal{N}+\mathcal{N})\sim P(\mu_Y,\Sigma_Y)$

$$\mu_Y=\mu_X+\mu_{X'}$$
$$\Sigma_{YY}=\Sigma_X+\Sigma_{X'}$$

$P(\mathcal{N}\mathcal{N})\sim P(\mu_Y,\Sigma_Y)$

$$\Sigma_{YY}=(\Sigma_{XX}^{-1}+\Sigma_{X'X'}^{-1})^{-1}$$
$$\mu_Y=\Sigma_{YY}\Sigma_{XX}^{-1}\mu_X+\Sigma_{YY}\Sigma_{X'X'}^{-1}\mu_{X'}$$

entropy: $ln(\sigma\sqrt{2\pi e})$ (max in all distribution at μ,Σ)

- KL-Divergence:

$$D_{KL}(p||q)=\int p(x)log\frac{p(x)}{q(x)}dx=H(p|q)-H(p)$$

- $D_{KL}(p||q)$(backward) : mode averaging
- $D_{KL}(q||p)$(forward) : mode seeking
  $p\sim\mathcal{N}(0,diag(\sigma_1^2,\sigma_2^2))\Rightarrow\sigma_q^2=\frac{2}{\sigma_1^{-2}+\sigma_2^{-2}}$
  $D_{KL}(q||p)$ is well defined if q is a subset of p
- $q\sim\mathcal{N}(\mathbb{E}(p),Var(p))\Rightarrow H(p|q)=H(q)$

- Bayesian Linear Regression:

$$y=w^\top x+\epsilon\quad \epsilon\sim\mathcal{N}(0,\sigma_n^2)\ w\sim\mathcal{N}(0,\sigma_p^2)$$
$$P(w|Y,X)\sim\mathcal{N}(\mu,\Sigma)$$
$$P(y_*\mid X,Y,x_*)=\mathcal{N}(\bar{\mu}^\top x_*,\ x_*^\top\bar{\Sigma}x_*+\sigma_n^2)$$
$$\bar{\mu}=\frac{1}{\sigma_n^2}\bar{\Sigma}X^\top Y\quad \bar{\Sigma}=\left(\frac{1}{\sigma_n^2}X^\top X+\frac{1}{\sigma_p^2}I\right)^{-1}$$

uncertainty: aleartoric(rand) + epistemic(know)

Online

$$X_{new}^\top X_{new}=X^\top X+x_{t+1}x_{t+1}^\top$$
$$X_{new}^\top Y_{new}=X^\top Y+y_{t+1}x_{t+1}$$

Fast : $\mathcal{O}(d^3)\to\mathcal{O}(d^2)$

$$(A+xx^\top)^{-1}=A^{-1}-\frac{(A^{-1}x)(A^{-1}x)^\top}{1+x^\top A^{-1}x}$$
$$(X_{new}^\top X_{new}+\sigma_n^2I)^{-1}=(\underbrace{X^\top X+\sigma_n^2}_{A}+x_{t+1}x_{t+1}^\top)^{-1}$$

Bayesian Logistic Regression

- posterior is not gaussian and not closed
- posterior log-density is convex

GP & Kalman Filter

- Gaussian Process

$$y=f(x)+\epsilon\quad \epsilon\sim\mathcal{N}(0,\sigma_n^2)$$
$$P(f|X,Y,x^*)\sim GP(f;\mu',k')$$
$$\mu'(x^*)=\mu(x^*)+K_{x^*X}(K_{XX}+\sigma_n^2I)^{-1}Y$$
$$k'(x^*)=K_{x^*x^*}-K_{x^*X}(K_{XX}+\sigma_n^2I)^{-1}K_{x^*X}^\top$$

$\mathcal{O}(n^3)$ for the inverse operation
prediction in closed form

Kernel

RBF
$$k(u,v)=\sigma_F^2exp\left(-\frac{(u-v)^2}{2l^2}\right)$$

l: length scale control the distance of data
σ_F : output scale control the magnitude

Kernel properties

- **Stationary**: translation invariant, $k(x,x')=k(x-x')$.
- **Isotropic/Stationary**: depends only on the distance, $k(x,x')=k(\|x-x'\|)$.
- **Isotropic**: rotation invariant, $k(Rx,Rx')=k(x,x')$.

Kernel conditions

$$K_{ij}=k(x_i,x_j)$$

- symmetry: $k(x,x')=k(x',x)$
- positive semi-definiteness: for any finite set $\{x_i\}_{i=1}^n$ and any $z\in\mathbb{R}^n$, $z^\top Kz\geq 0$ (ensures variances ≥ 0)

Kernel composition rules

- $k_1,k_2: X\times X\to\mathbb{R}$
- $k(x,x')=k_1(x,x')+k_2(x,x')$
- $k(x,x')=c\,k_1(x,x'),\quad c>0$
- $k(x,x')=k_1(x,x')\,k_2(x,x')$
- $k(x,x')=f\big(k_1(x,x')\big)$
- $k(x,x')=\sum_{i=1}^n\phi_i(x)\,\phi_i(x')$ (feature maps are valid)

Bivariate covariance

$$\mathrm{Cov}\begin{pmatrix}X\\Y\end{pmatrix}=\begin{bmatrix}\mathrm{Var}(X)&\mathrm{Cov}(X,Y)\\\mathrm{Cov}(X,Y)&\mathrm{Var}(Y)\end{bmatrix}$$

where $\mathrm{Cov}(X,Y)=k(x,y)$ (kernel function)

- Hyperparameter optimization

$$\log p(y\mid X,\theta)=-\tfrac{1}{2}y^\top K_y^{-1}y-\tfrac{1}{2}\log|K_y|-\tfrac{n}{2}\log(2\pi),$$
$$K_y:=K_{f,\theta}+\sigma_n^2I,\ \alpha:=K_y^{-1}y$$
$$\frac{\partial}{\partial\theta_j}\log p(y\mid X,\theta)=\tfrac{1}{2}\mathrm{tr}\Big((\alpha\alpha^\top-K_y^{-1})\frac{\partial K_y}{\partial\theta_j}\Big)$$

GP & Kalman Filter

Fast

- $k(x,x')=\phi(x)\phi(x)^\top$ $\mathcal{O}(n^3)\to\mathcal{O}(nm^2+m^3)$
- fourier features: $k(x,x')\approx k(x-x')$
  $=\int_{\mathbb{R}^d}p(\omega)e^{j\omega^top(x-x')}dw$ (stationary)
  Bochner Theorem:$p(\omega)\geq 0\Rightarrow k\geq 0$
- inducing points: $\mathcal{O}(n^3)$
  cubic inducing points, linear point
  SoR: subset of regressor (zero)
  FITC: (diag)

- Kalman Filter

$$x_{t+1}=F_tx_t+\Sigma_{x,t}\quad y_t=H_tx_t+\Sigma_y$$

$P(x_{t+1}|x_t)\sim\mathcal{N}(Fx_t,\Sigma_{x,t})$ $P(y_t|x_t)\sim\mathcal{N}(H_tx_t,\Sigma_y)$

predict

$$\hat{x}_{t+1}=F_tx_t\quad \hat{\Sigma}_{x,t+1}=H_t\Sigma_{x,t}H_t^\top+\Sigma_{d_t}$$
$$P(x_t|y_{1:t})\sim\mathcal{N}(\mu_t,\Sigma_{d_t}^2)$$

correct

$$K_{t+1}=\hat{\Sigma}_{x,t}H_t^\top(H_t\hat{\Sigma}_{x,t}H_t^\top+\Sigma_y)^{-1}$$
$$x_{t+1}=\hat{x}_{t+1}+K_{t+1}(y_{t+1}-H_t\hat{x}_{t+1})$$
$$\Sigma_{x,t+1}=(I-K_{t+1}H_t)\hat{\Sigma}_{x,t}$$

kalman gain:K_{t+1}, can be computed offline

Linear Dynamic System

$$x_{k+1}\sim(ax_k,\sigma^2),x_0=0\to MLE(a)=\frac{\sum_{i=0}^{n-1}x_ix^{i+1}}{\sum_{i=0}^{n-1}x_i^2}$$

if $a\sim\mathcal{N}(0,\frac{\sigma^2}{\lambda})$, $Var(a|x_0,\cdots,x_n)=\frac{\sigma^2}{\lambda+\sum_{i=0}^{n-1}x_i^2}$

Prior / Likelihood

$$p(\theta)\,,\qquad p(\mathcal{D}\mid\theta)=\prod_{i=1}^n p(y_i\mid x_i,\theta)\quad\text{(i.i.d.)}$$

Bayesian update

$$p(\theta\mid\mathcal{D})=\frac{p(\mathcal{D}\mid\theta)\,p(\theta)}{p(\mathcal{D})}\quad p(\mathcal{D})=\int p(\mathcal{D}\mid\theta)\,p(\theta)\,d\theta$$

Posterior predictive (prediction)

$$p(y_*\mid x_*,\mathcal{D})=\int p(y_*\mid x_*,\theta)\,p(\theta\mid\mathcal{D})\,d\theta$$

Probability laws(continuous case)

$$Probability: P(A)=\int P(A\mid B=b)\,p_B(b)\,db$$
$$Expectation: \mathbb{E}[X]=\mathbb{E}\big[\mathbb{E}[X\mid Y]\big]$$
$$Var: \mathrm{Var}(X)=\mathbb{E}\big[\mathrm{Var}(X\mid Y)\big]+\mathrm{Var}\,\big(\mathbb{E}[X\mid Y]\big)$$

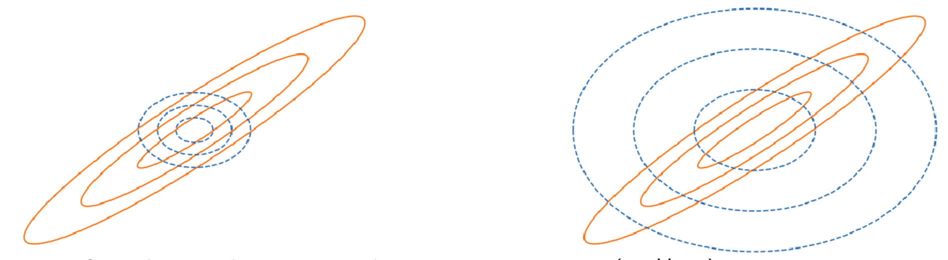
Approximation & MCMC

- Approximation

Laplacian approximation $q(\theta)\sim\mathcal{N}(\hat{\theta},\Lambda^{-1})$

$$\hat{\theta}=\underset{\theta}{argmax}\ p(\theta|y)\quad \Lambda=-\nabla\nabla log(p(\hat{\theta}|y))$$

- not skewed
- only one modal
- no previous knowledge



- Left: backward KL $D_{KL}(q\|p)$
- Right: forward KL $D_{KL}(p\|q)$

Evidence Lower Bound(ELBO)

$$Q^*=\underset{Q\in\mathcal{Q}}{argmin}\ D_{KL}(Q(Z)\|P(Z|X))$$
$$=\underset{Q\in\mathcal{Q}}{argmax}\ \mathbb{E}_{Z\sim Q(Z)}[log(P(X|Z))]$$
$$-D_{KL}(Q(Z)||P(Z))=\mathbb{E}_qp-D_{KL}(p\|p)$$

MLE $\hat{\theta}=\underset{\theta}{argmax}\ P(X_{1:n}|\theta)$

sometimes add l_2 norm which is $\|\cdot-\mu\|_2^2$ for vector and $\|\cdot-\mu\|_F^2$ for matrix and $(\cdot-\mu)^2$ for scalar

MAP $\hat{\theta}=\underset{\theta}{argmax}(X_{1:n}|\theta,X')$

reparameterization tricks

- random variable continous
- reduce variance by introducing bias
- automatic differentiation

- Markov Chain Monte Carlo

$$\pi=\pi P$$

π is the stationary state, P is the transition matrix

ergodic:$\exists t\in\mathbb{N}\to(P)^t>0$

- MCMC used for any distribution
- $P(x|x')=P(x'|x)\to$ uniform distribution

Metropolis-Hastings (MH) Algorithm:

given proposal transition $R(x'|x)$ and unnormalized stationary distribution $Q(x)$

accept rate $\alpha=min\{1,\frac{Q(x')R(x|x')}{Q(x)R(x'|x)}\}$ from x to x'

Gibbs Sampling : random choose dimension

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

Stochastic Gradient Langevin Dynamics(SGLD)

$$\Delta\theta=-\eta(\nabla log\ p(\theta_t)+\frac{N}{n}\sum_{j=1}^n\nabla logp(y_{i_j}|\theta_t,x_{i_j}))+\epsilon_t$$
$$=-\eta(\theta_t+\nabla_\theta log\ L(\theta_t))+\epsilon_t$$

$\epsilon_t\sim\mathcal{N}(0,2\eta I)$

$L(\theta_t)$ is the likelihood

can guarantee convergence, need burn in step

ETHz PAI Cheat Sheet

Bayesian DL & Optimization

- Bayesian DeepLearning

$$P(y|x,\theta) \sim \mathcal{N}(y|f_\mu(x;\theta_\mu), exp(f_\sigma(x;\theta_\sigma)))$$

train

$$\hat\theta = \underset{\theta}{argmin} - lnP(\theta) - \sum_{i=1}^{n} lnP(y_i|x_i,\theta)$$

$$= \underset{\theta}{argmin}\ \lambda||\theta||_2^2 + \frac{1}{2}\sum_{i=1}^{n}\left[\frac{||y_i-\mu(x_i;\theta_\mu)||^2}{\sigma(x_i;\theta_\sigma)^2}\right.$$

$$\left. + ln\ \sigma(x_i;\theta_\sigma)^2\right]$$

predict $P(y'|x',X,Y) = \mathbb{E}_{\theta\sim q}[P(y'|x',\theta)]$

$$Var[y'|x',X,Y] = \mathbb{E}_{\theta\sim q}[Var[y'|x',\theta]] + Var[\mathbb{E}_{\theta\sim q}[y'|x',\theta]]$$

- Aleatoric uncertainty(random)
- Epistemic uncertainty(knowledge)

- MAP for BNN is not closed, unless likelihood and prior are $\mathcal{N}$
- MLE for BNN is closed
- $Var(C\theta|\theta) = 0 \quad \theta \sim p$

approximate inference for BNN: 1.SGLD, 2.Dropout, 3.Ensemble, 4.black-box stochastic variational inference

Stochastic Weight Averaging-Gaussian(SWAG)

$$\mu_{SWA} = \frac{1}{T}\sum_{1}^{T}\theta_i$$

$$\Sigma_{SWA} = \frac{1}{T-1}\sum_{1}^{T}(\theta_i-\mu_{SWA})(\theta_i-\mu_{SWA})^\top$$

Dropout

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

- dropout is also applied in prediction
- dropout can be seen as varational inference

- Bayesian Optimization

Mutual Information

$$I(s) = H(f) - H(f|y_s) = \frac{1}{2}log|I+\sigma_n^{-2}K_s|$$

$$F(S_T) \geq (1-\frac{1}{e})\underset{S\subseteq D,|S|\leq T}{max} F(S)$$

regret

$$R_T = \sum_{t=1}^{T}(max_{x\in D}f(x) - f(x_t))$$

- sublinear(optimal) if $\frac{R_T}{T} \to 0$ $\lim_{t\to\infty} f(x_t) \to f(x^*)$
- $R_T^A \leq R_T^B$ cannot tell anything
- $\forall t\ R_t^A \leq R_t^B$ menas A is better
- $R\uparrow\to$ more exploration

Bayesian Optimization

- Bayesian Optimization

Uncertainty Sampling

$$x_t = \underset{x\in D}{argmax}\frac{\sigma_e^2(x)}{\sigma_a(x)}$$

- ✓ max info gain in homoscedastic noise case
- ✗ max info gain in heterodastic noise
- ✓ $\lim_{t\to\infty} \hat x_t = \underset{x\in D}{argmax}\ \mu_t(x), f(\hat x) \to f(x^*)$
- ✗ $\lim_{t\to\infty} f(x_t) \to f(x^*)$

Upper Confidence Sampling(UCB)

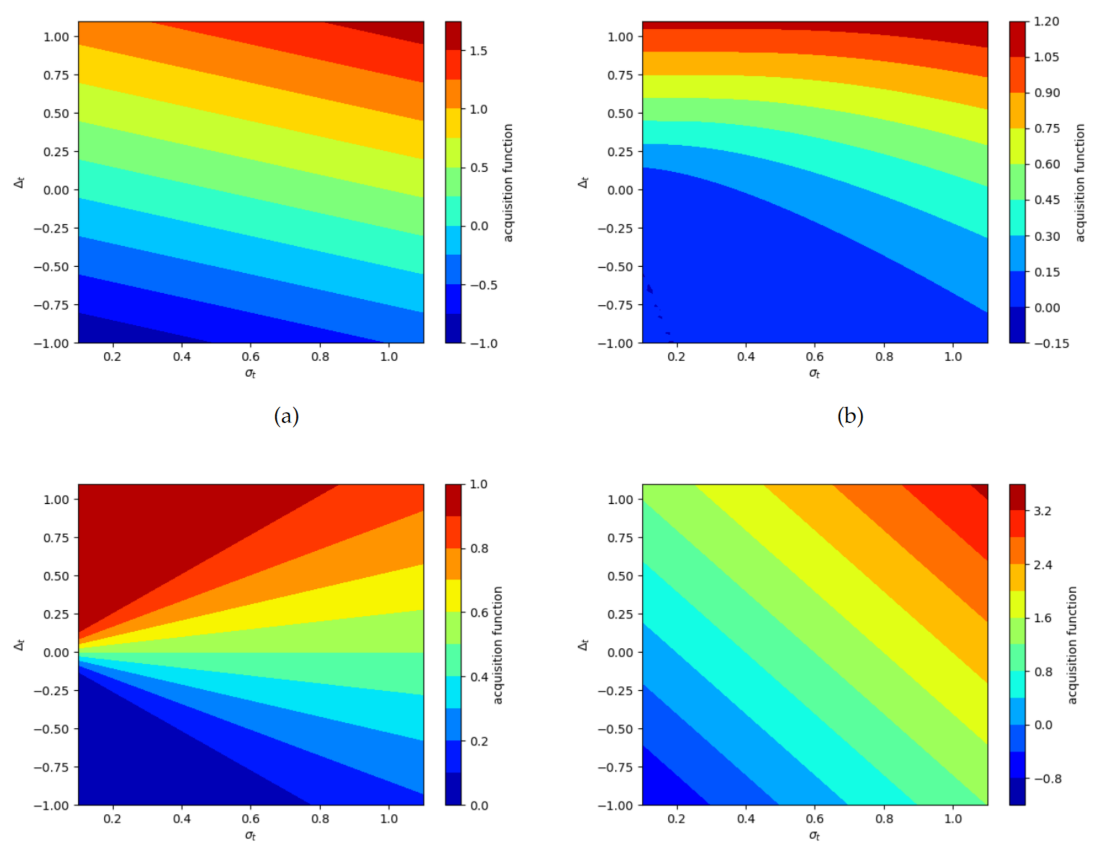
$$a = \mu_t(x) + \beta\sigma_t(x)$$

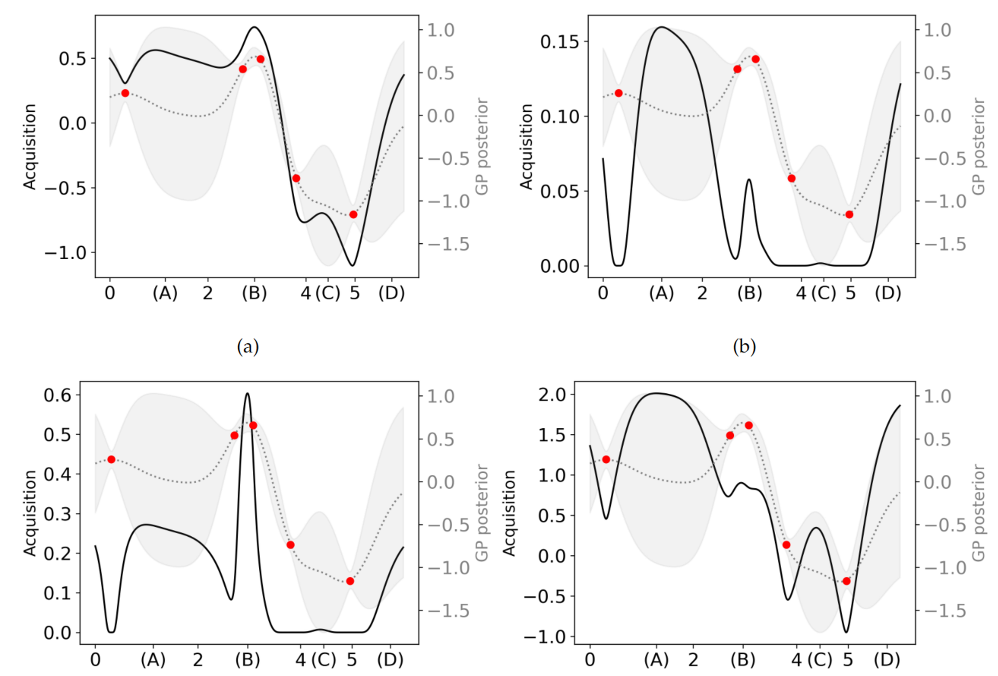
Probability of Improvement(PI)

$$a = \Phi\left(\frac{\mu_t(x)-f^*}{\sigma_t(x)}\right)$$

Expected Improvement(EI)

$$a = (\mu_t(x)-f^*)\Phi\left(\frac{\mu_t(x)-f^*}{\sigma_t(x)}\right) + \sigma_t(x)\phi\left(\frac{\mu_t(x)-f^*}{\sigma_t(x)}\right)$$





UCB($\beta=0.5$)	EI
PI	UCB($\beta=2$)

$$acquisition \uparrow \Longleftrightarrow exploitation$$

$$acquisition \downarrow \Longleftrightarrow exploration$$

Reinforcement Learning

Bellman Theorem

$$V^*(x) = max_a(r(x,a) + \gamma\sum_{x'}P(x'|x,a)V^*(x'))$$

Hoeffding Bound

$$P(|\mu - \frac{1}{n}\sum_{i=1}^{n}Z_i| > \varepsilon) \leq 2exp(-\frac{2n\varepsilon^2}{C^2})$$

- cR do not change policy

- Model Based

Value Iteration

- guarantee converge to an ε optimal policy not the exact optimal policy
- polynomial time
- performance depend on the input

Policy Iteration

- monotonically improve the policy
- polynomial time, gaurantee converge

ϵ greedy

when random number $< \epsilon$ do the random action

Rmax method

set reward R and transition probability $P(x^*|x,a) = 1$ at first

- with probability $1-\sigma$,reach ε - optimal
- polynomial time in $|X|, |A|, T, \frac{1}{\varepsilon}, log(\frac{1}{\delta})$

- Model Free

Temporal Difference(TD) - Learning

$$\hat V^\pi(x) \leftarrow (1-\alpha_t)\hat V^\pi(x) + \alpha_t(r+\gamma\hat V^\pi(x'))$$

on-policy

Theorem:$\sum_t\alpha_t = \infty, \sum_t\alpha_t^2 < \infty \Rightarrow P(\hat V^\pi \to V^\pi) = 1$

Q-Learning

off-policy

$$\hat Q^* \leftarrow (1-\alpha_t)\hat Q^*(x,a) + \alpha_t(r+\gamma\underset{a'}{max}\hat Q^*(x',a'))$$

init : $\hat Q^*(x,a) = \frac{R_{max}}{1-\gamma}\Pi_{t=1}^{T_{init}}(1-\alpha_t)^{-1}$

Theorem: $\sum_t\alpha_t = \infty, \sum_t\alpha_t^2 < \infty \Rightarrow P(\hat Q^* \to Q^*) = 1$

- with probability $1-\sigma$, R max will reach an ε - optimal
- polynomial time in $|X|, |A|, T, \frac{1}{\varepsilon}, log(\frac{1}{\delta})$
- decay learning rate guarantee convergence

Deep RL

- Model Free

Policy Search

REINFORCE

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim p_\theta(\tau)}[\sum_{t=0}^{\top}\gamma^t(\sum_{t'=t}^{\top}\gamma^{t'-t}r_{t'})\nabla_\theta ln\pi_\theta(a_t|x_t)]$$

$$\theta \leftarrow \theta + \eta_t\nabla_\theta J(\theta)$$

Actor-Critic

$$\nabla_{\theta_\pi}J(\theta_\pi) = \mathbb{E}_{(x,a)\sim\pi_{\theta_\pi}}[Q_{\theta_Q}(x,a)\nabla_{\theta_\pi}ln\pi_{\theta_\pi}(a|x)]$$

$$\theta_\pi \leftarrow \theta_\pi + \eta_t\nabla_{\theta_\pi}J(\theta_\pi)$$

$$\theta_Q \leftarrow \theta_Q - \eta_t(Q_{\theta_Q}(x,a) - r - \gamma Q_{\theta_Q}(x',\pi_{\theta_\pi}(x')))\nabla_{\theta_Q}Q_{\theta_Q}(x,a)$$

- use a baseline to reduce variance in the gradient estimates

Proximal Policy Optimization(PPO)

$$L_{\theta_k}(\theta_k) = \mathbb{E}_{\tau\sim\pi_k}\sum_{t=0}^{\infty}\left[\frac{\pi_\theta(a|x)}{\pi_{\theta_k}(a|x)}(r+\gamma Q^{\pi_{\theta_k}}(x',a) - Q^{\pi_{\theta_k}}(x,a))\right]$$

$$\theta_k \leftarrow \theta_k - \eta_t\nabla_{\theta_k}L_{\theta_k}(\theta_k)$$

Deep Deterministic Policy Gradients(DDPG)

- randomly add noise ensure sufficient exploration

on-policy	REINFORCE, Actor-Critic, PPO
off-policy	DDPG, TD3, SAC

- Model Based

reduce sample complexity

Random Shooting Method

Monte Carlo Tree Search

PETS

- Other

Bernoulli distribution

$$Bernoulli(x;p) = p^x(1-p)^{1-x}$$

$$\mathbb{E}[Bernoulli(x;p)] = p$$

$$\mathbb{V}ar[Bernoulli(x;p)] = p(1-p)$$

Poisson distribution

$$Pr(x;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$\mathbb{E}[Pr(x;\lambda)] = \lambda$$

$$\mathbb{V}ar[Pr(x;\lambda)] = \lambda$$