Disclaimer

This document is an exam summary that follows the slides of the *Probabilistic Artificial Intelligence* lecture at ETH Zurich. The contribution to this is a short summary that includes the most important concepts, formulas and algorithms. This summary was created during the fall semester 2020. Due to updates to the syllabus content, some material may no longer be relevant for future versions of the lecture. This work is published as CC BY-NC-SA.

I do not guarantee correctness or completeness, nor is this document endorsed by the lecturers. Feel free to point out any erratas. For the full LaTeX source code, consider `github.com/ymerkli/eth-summaries`.

Terms and Acronyms

Consult the following list of acronyms in case any of them are unclear:

- BALD: Bayesian Active Learning by Disagreement
- BE: Bellman Equation
- BLinR: Bayesian Linear Regression
- BLogR: Bayesian Logistic Regression
- BNN: Bayesian Neural Network
- BbB: Bayes by Backprop
- CDF: Cumulative Distribution Function
- CoV: Change of Variable
- DBE: Detailed Balance Equation
- DDPG: Deep Deterministic Policy Gradient
- EI: Expected Improvement
- FA: Function Approximation
- FITC: Fully Independent Training Conditional
- GP-UCB: Gaussian Process Upper Confidence Bound
- GP: Gaussian Process
- GS: Gibbs Sampling
- HMM: Hidden Markov Model
- KL: Kullback–Leibler divergence
- MAP: Maximum A Posteriori
- MC: Markov Chain
- MCMC: Markov Chain Monte Carlo
- MDP: Markov Decision Process
- MLE: Maximum Likelihood Estimation
- MPC: Model Predictive Control
- PDF: Probability Density Function
- PETS: Probabilistic Ensembles with Trajectory Sampling
- PI: Policy Iteration
- POMDP: Partially observable Markov decision process
- PSD: Positive Semi-Definite
- RM: Robbins Monro
- RV: Random Variable
- SG-HMC: Stochastic Gradient Hamiltonian Monte Carlo
- SGD: Stochastic Gradient Descent
- SGLD: Stochastic gradient Langevin dynamics
- TD-Learning: Temporal Difference Learning
- VI: Variational Inference/ Value Iteration

Basics
Product: $P(X,Y)=P(X|Y)P(Y)=P(Y|X)P(X)$
Chain: $P(X_1,X_2,...,X_n)=P(X_{1:n})=$
$=P(X_1)P(X_2|X_1)P(X_3|X_{1:2})...P(X_n|X_{1:n-1})$
Sum: $P(X_{1:n})=\sum_y P(X_{1:n},Y=y)=$
$\sum_y P(X_{1:n}|Y=y)P(Y=y)=\int_y P(X_{1:n}|Y=y)P(Y=y)dy$
Bayes: $P(X|Y)=\frac{P(X,Y)}{P(Y)}=\frac{P(Y|X)P(X)}{P(Y)}$
X, Y indep.: $P(X|Y)=P(X), P(X,Y)=P(X)P(Y)$
Expec: $\mathbb{E}_x[f(X)]=\int f(x)p(x)dx=\sum_x f(x)p(x)$
Lin Exp: $\mathbb{E}_{x,y}[aX+bY]=a\mathbb{E}_x[X]+b\mathbb{E}_y[Y]$
Variance: $Var[X]=\mathbb{E}[(X-\mu_X)^2]=\mathbb{E}[X^2]-\mathbb{E}[X]^2$
$Var(X+Y)=Var(X)+Var(Y)+2Cov(X,Y)$
Covariance: $Cov(X,Y)=\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]$
CoV: $Y=g(X), f_Y(y)=f_X(g^{-1}(y))\cdot|\frac{d}{dy}g^{-1}(y)|$
Gauss: $\mathcal{N}=\frac{1}{\sqrt{(2\pi)^d|\Sigma|}}exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$
CDF: $\Phi(u;\mu,\sigma^2)=\int_{-\infty}^u \mathcal{N}(y;\mu,\sigma^2)dy=\Phi(\frac{u-\mu}{\sqrt{\sigma^2}};0,1)$;
Multivar. Gauss: $X_V=[X_1,..,X_d]\sim\mathcal{N}(\mu_V,\Sigma_{VV})$, index sets $A=\{i_1,...,i_k\}, B=\{j_1,...,j_m\}, A\cap B=\emptyset$
Marginal: $X_A=[X_{i_1}..X_{i_k}]\sim\mathcal{N}(\mu_A,\Sigma_{AA})$ with $\mu_A=[\mu_{i_1},..,\mu_{i_k}], \Sigma_{AA}^{(m,n)}=\sigma_{i_m,i_n}=\mathbb{E}[(x_{i_m}-\mu_{i_m})(x_{i_n}-\mu_{i_n})]$
Conditional: $P(X_A|X_B=x_B)=\mathcal{N}(\mu_{A|B},\Sigma_{A|B})$ with $\mu_{A|B}=\mu_A+\Sigma_{AB}\Sigma_{BB}^{-1}(x_B-\mu_B)$ and $\Sigma_{A|B}=\Sigma_{AA}-\Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}$
$Y=MX_A, M\in\mathbb{R}^{m\times d}$, $Y\sim\mathcal{N}(M\mu_A, M\Sigma_{AA}M^T)$
$Y=X_A+X_B$, $Y\sim\mathcal{N}(\mu_A+\mu_B,\Sigma_{AA}+\Sigma_{BB})$
KL: $KL(p||q)=\mathbb{E}_p[log\frac{p(x)}{q(x)}]=\sum_{x\in X}p(x)\cdot log\frac{p(x)}{q(x)}$ $=\int p(x)log\frac{p(x)}{q(x)}dx\geq 0$, $p=q: KL(p||q)=0$
Entropy: $H(q)=\mathbb{E}_q[-\log q(\theta)]=-\int q(\theta)\log q(\theta)d\theta$ $-\sum_\theta q(\theta)\log q(\theta)$; $H(\prod q_i(\theta_i))=\sum_i H(q_i)$;
$H(N(\mu,\Sigma))=\frac{1}{2}ln|2\pi e\Sigma|$; $H(p,q)=H(p)+H(q|p)$
$H(S|T)\geq H(S|T,U)$ 'information never hurts'
Orth: A: $A^{-1}=A^T, AA^T=A^TA=||A||_2^2=I$
$det(A)\in\{+1,-1\},(A^{-1})^T=(A^T)^{-1}, rank(A)=n$
Inv: $A^{-1}=\begin{bmatrix}a&b\\c&d\end{bmatrix}^{-1}=\frac{1}{ad-bc}\begin{bmatrix}d&-b\\-c&a\end{bmatrix}$;
Deriv: $(fg)'=f'g+fg'$; $(f/g)'=(f'g-fg')/g^2$
$f(g(x))'=f'(g(x))g'(x)$; $\log(x)'=1/x$
Convex: g(x) is convex $\Leftrightarrow x_1,x_2\in\mathbb{R},\lambda\in[0,1]$: $g''(x)>0$; $g(\lambda x_1+(1-\lambda)x_2)\leq\lambda g(x_1)+(1-\lambda)g(x_2)$
Jensen inequality: g convex: $g(E[X])\leq E[g(X)]$
g concave (e.g. log): $g(E[X])\geq E[g(X)]$
Bayesian Learning: Prior $p(\theta)$;
Likelihood $p(y_{1:n}|x_{1:n},\theta)=\prod_{i=1}^n p(y_i|x_i,\theta)$;
Posterior $p(\theta|x_{1:n},y_{1:n})=\frac{1}{Z}p(\theta)\prod_{i=1}^n p(y_i|x_i,\theta)$;
where $Z=\int p(\theta)\prod_{i=1}^n p(y_i|x_i,\theta)d\theta$; Prediction:
$p(y^*|x^*,x_{1:n},y_{1:n})=\int p(y^*|x^*,\theta)p(\theta|x_{1:n},y_{1:n})d\theta$

BLinR $f^*=\mathbf{w}^T\mathbf{x}^*, y^*=f^*+\epsilon,\epsilon\sim\mathcal{N}(0,\sigma_y^2)$
$p(\mathbf{w})=\mathcal{N}(0,\sigma_w^2\mathbf{I}), p(y_i|\mathbf{x_i},\mathbf{w},\sigma_y)=\mathcal{N}(y_i;\mathbf{w}^T\mathbf{x_i},\sigma_y^2)$
$p(\mathbf{w}|\mathbf{X},\mathbf{y})=\mathcal{N}(\mathbf{w};\overline{\mu},\overline{\Sigma})$, $\overline{\Sigma}=(\sigma_y^{-2}\mathbf{X}^T\mathbf{X}+\sigma_w^{-2}\mathbf{I})^{-1}$,
$\overline{\mu}=\sigma_y^{-2}\overline{\Sigma}\mathbf{X}^T\mathbf{y}$; $p(f^*|\mathbf{X},\mathbf{y},\mathbf{x}^*)=\mathcal{N}(\mathbf{x}^{*T}\overline{\mu},\mathbf{x}^{*T}\overline{\Sigma}\mathbf{x}^*)$;
$p(y^*|\mathbf{X},\mathbf{y},\mathbf{x}^*)=\mathcal{N}(\mathbf{x}^{*T}\overline{\mu},\mathbf{x}^{*T}\overline{\Sigma}\mathbf{x}^*+\sigma_y^2)$
Epistemic: uncertainty about model due to lack of data. **Aleatoric**: Irreducible noise
Recursive updates: $\mathbf{X}_{t+1}^T\mathbf{X}_{t+1}=\mathbf{X}_t^T\mathbf{X}_t+x_{t+1}x_{t+1}^T$
$\mathbf{X}_{t+1}^T y_{t+1}=\mathbf{X}_t^T y_t+y_{t+1}x_{t+1}$
BLogR $p(y_i|x_i,\theta)=\sigma(y_iw^Tx_i), \sigma(a)=\frac{1}{1+e^{-a}}$
Kalman Filter $X_{t+1}\perp X_{1:t-1}|X_t, Y_t\perp Y_{1:t-1},X_{1:t-1}|X_t$
State X_t, Observation Y_t, Prior $P(X_1)\sim\mathcal{N}(\mu,\Sigma)$
Motion model: $P(\mathbf{X}_{t+1}|\mathbf{X}_t)=\mathcal{N}(x_{t+1};\mathbf{F}X_t,\Sigma_x)$,
$\mathbf{X}_{t+1}=\mathbf{F}\mathbf{X}_t+\epsilon_t, \epsilon_t\sim\mathcal{N}(0,\Sigma_x)$
Sensor model: $P(\mathbf{Y}_t|\mathbf{X}_t)=\mathcal{N}(y_t;HX_t,\Sigma_y)$,
$\mathbf{Y}_t=\mathbf{H}\mathbf{X}_t+\eta_t,\eta_t\sim\mathcal{N}(0,\Sigma_y)$
Kalman update: $\mu_{t+1}=\mathbf{F}\mu_t+\mathbf{K}_{t+1}(\mathbf{y}_{t+1}-\mathbf{HF}\mu_t)$
$\Sigma_{t+1}=(\mathbf{I}-\mathbf{K}_{t+1}\mathbf{H})(\mathbf{F}\Sigma_t\mathbf{F}^T+\Sigma_x)$
Kalman gain: $\mathbf{K}_{t+1}=(\mathbf{F}\Sigma_t\mathbf{F}^T+\Sigma_x)\cdot$ $\cdot\mathbf{H}^T(\mathbf{H}(\mathbf{F}\Sigma_t\mathbf{F}^T+\Sigma_x)\mathbf{H}^T+\Sigma_y)^{-1}$
Bayesian Filtering in KFs Keep track of state X_t using rec. formula. Start $P(X_1)=\mathcal{N}(\mu,\Sigma)$.
At time t: assume we have $P(X_t|y_{1:t-1})$
Conditioning: $P(X_t|y_{1:t})=\frac{1}{Z}P(y_t|X_t)P(X_t|y_{1:t-1})$
Prediction: $P(X_{t+1}|y_{1:t})=\int P(X_{t+1}|x_t)P(x_t|y_{1:t})dx_t$
Gaussian Processes Gaussian distr. over functions $f\sim GP(\mu(x),K(x)$ (∞-dim Gaussian).
Infinite set of RVs X s.t. $\forall A\subseteq X, A=\{x_1,..,x_m\}$ it holds $Y_A=[Y_{x_1},..,Y_{x_m}]\sim\mathcal{N}(\mu_A,K_{AA})$ where $K_{AA}^{(ij)}=k(x_i,x_j)$ and $\mu_A^{(i)}=\mu(x_i)$ with covariance function $k(\cdot,\cdot)$, mean function $\mu(\cdot)$
Covariance k: symmetric, PSD, kernel composition rules hold, stationary: $k(x,x')=k(x-x')$, isotropic: if $k(x,x')=k(||x-x'||_2)$.
GP Prediction $p(f)=GP(f;\mu(x),k(x,x'))$, observe $y_i=f(x_i)+\epsilon_i$, $\epsilon_i\sim\mathcal{N}(0,\sigma^2)$, $A=\{x_1,..,x_m\}$.
Common convention: prior mean $\mu(x)=0$
Then $p(f|x_{1:m},y_{1:m})=GP(f;\mu',k')$ where
$\mu'(x)=\mu(x)+\mathbf{k}_{x,A}(\mathbf{K}_{AA}+\sigma^2\mathbf{I})^{-1}(\mathbf{y}_A-\mu_A)$
$k'(x,x')=k(x,x')-\mathbf{k}_{x,A}(\mathbf{K}_{AA}+\sigma^2\mathbf{I})^{-1}\mathbf{k}_{x',A}^T$
$k_{x,A}=[k(x,x_1),...,k(x,x_m)]$
Predictive posterior: $p(y^*|x_{1:m},y_{1:m},x^*)=\mathcal{N}(\mu_y^*,\sigma_y^{2*}), \mu_y^*=\mu'(x^*), \sigma_y^{2*}=\sigma^2+k'(x^*,x^*)$
Forward sampling GP: Chain rule on $P(f_1,..,f_n)$, iteratively sample univariate Gauss
Model selection: max. marginal likelihood
$\hat\theta=amax_\theta p(y|X,\theta)=amax_\theta\int p(y|X,f)p(f|\theta)df$
Fast GPs: GP prediction has cost $\mathcal{O}(|A|^3)$

- Local: distance decaying kernel (e.g. RBF), only condition on points x' where $|k(x,x')|>\tau$
- k approx: $k(x,x')\approx\phi(x)^T\phi(x')$, then do BLR
- RFF: Stationary kernel has Fourier transf.: $k(x,x')=\int_{\mathbb{R}^d}p(\omega)e^{j\omega^T(x-x')}d\omega=$
$\mathbb{E}_{\omega,b}[z_{w,b}(x)z_{w,b}(x')]\approx\frac{1}{m}\sum_i z_{w^{(i)},b^{(i)}}(x)z_{w^{(i)},b^{(i)}}(x')$,
$\omega\sim p(\omega), b\sim\mathcal{U}[0,2\pi], z_{\omega,b}(x)=\sqrt{2}\cos(\omega^Tx+b)$
$\to k(x,x')\approx\phi(x)^T\phi(x')$ $(\phi_i(x)=\frac{1}{\sqrt{m}}z_{w^{(i)},b^{(i)}}(x))$
- **Inducing Points Methods**: Summarize data via values of f at inducing points $\mathbf{u}=[u_1,..,u_m]$.
$p(f^*,f)=\int p(f^*,f,u)du=\int p(f^*,f|u)p(u)du$
$p(f^*,f)\approx q(f^*,f)=\int q(f^*|u)q(f|u)p(u)du$
with $p(f|u)=\mathcal{N}(K_{f,u}K_{u,u}^{-1}u,K_{f,f}-Q_{f,f})$,
$p(f^*|u)=\mathcal{N}(K_{f^*,u}K_{u,u}^{-1}u,K_{f^*,f^*}-Q_{f^*,f^*})$,
and $Q_{a,b}=K_{a,u}K_{u,u}^{-1}K_{u,b}$, $p(\mathbf{u})\sim\mathcal{N}(0,K_{u,u})$
Subset of Regressors: assume $K_{f,f}-Q_{f,f}=0$, replace $p(f|u)$ by $q_{SoR}(f|u)=\mathcal{N}(K_{f,u}K_{u,u}^{-1}u,0)$ resulting model is degenerate GP with covariance function $k_{SoR}(x,x')=k(x,u)K_{u,u}^{-1}k(u,x')$
FITC: Assume $f_i\perp\!\!\!\perp f_j|u,\forall i\neq j$
$q_{FITC}(f|u)=\mathcal{N}(K_{f,u}K_{u,u}^{-1}u,diag(K_{f,f}-Q_{f,f}))$
Laplace Approx $p(w|(x,y)_{1:n})\approx q_\lambda(\theta)=\mathcal{N}(\hat\theta,\Lambda^{-1})$
$\hat\theta=\arg\max_\theta p(\theta|y),\Lambda=-\nabla\nabla\log p(\hat\theta|y)$
Predict: $p(y^*|x^*,x_{1:n},y_{1:n})\approx\int p(y^*|f^*)q(f^*)df^*$, with $q(f^*)=\int p(f^*|\theta)q_\lambda(\theta)d\theta$. LA first greedily fits mode, then matches curvature (over-conf.).
Variational Inference $p(\theta|y)=\frac{1}{Z}p(\theta,y)\approx q_\lambda(\theta)$
$q^*_{bwd}\in\arg\min_{q\in Q}KL(q||p)$: $q\approx p$ where q large
$q^*_{fwd}\in\arg\min_{q\in Q}KL(p||q)$: $q\approx p$ where p large
$amin_q KL(q||p)=amax_q\mathbb{E}_{\theta\sim q}[\log p(\theta,y)]+H(q(\theta))$
$=amax_q\mathbb{E}_{\theta\sim q_\lambda(\theta)}[\log p(y|\theta)]-KL(q(\theta)||p(\theta))$
ELBO: $amax_q\mathbb{E}_{\theta\sim q_\lambda}[\log p(y|\theta)]-KL(q(\theta)||p(\theta))$
$\leq\log p(y)\to\nabla_\lambda L(\lambda)$ tricky due to $\theta\sim q_\lambda(\cdot)$
Reparametrization Trick: Suppose $\epsilon\sim\phi$, $\theta=g(\epsilon,\lambda)$. Then: $q(\theta|\lambda)=\phi(\epsilon)|\nabla_\epsilon g(\epsilon;\lambda)|^{-1}$ and $\mathbb{E}_{\theta\sim q_\lambda}[f(\theta)]=\mathbb{E}_{\epsilon\sim\phi}[f(g(\epsilon;\lambda))]$, which allows $\nabla_\lambda\mathbb{E}_{\theta\sim q_\lambda}[f(\theta)]=\mathbb{E}_{\epsilon\sim\phi}[\nabla_\lambda f(g(\epsilon;\lambda))]$
Markov Chains A stationary MC is a sequence of RVs $X_1,..,X_N$ with prior $P(X_1)$ and transition probability $P(X_{t+1}|X_t)$ independent of t. MC is ergodic if $\exists\ t<\infty$ s.t. every state is reachable from every state in *exactly* t steps.
Markovian Assumption: $X_{t+1}\perp\!\!\!\perp X_{1:t-1}|X_t\ \forall t$
Stationary Distribution: A stationary ergodic MC has a unique and positive stationary distr. $\pi(X)>0$ s.t. $\forall x: \lim_{N\to\infty}P(X_N=x)=\pi(x)$ and $\pi(X)$ is independent of prior $P(X_1)$.
Simulate MC via forward sampling (chain rule)

MCMC Approx pred. distr. $p(y^*|x^*,x_{1:n},y_{1:n})=\int p(y^*|x^*,\theta)p(\theta|(x,y)_{1:n})d\theta=\mathbb{E}_{\theta\sim p(\cdot|(x,y)_{1:n})}[f(\theta)]$
$\approx\frac{1}{m}\sum_{i=1}^m f(\theta^{(i)})$, sample $\theta^{(i)}\sim p(\theta|(x,y)_{1:n})$ from MC with stationary distribution $p(\theta|(x,y)_{1:n})$. **Hoeffding**: Assume $f\in[0,C]$:
$P(|\mathbb{E}_P[f(X)]-\frac{1}{N}\sum_{i=1}^N f(x_i)|>\epsilon)\leq 2exp(-2N\epsilon^2/C^2$
Given unnormalized distr. $Q(x)>0$, design MC s.t. $\pi(x)=\frac{1}{Z}Q(x)$. If MC satisfies **detailed balance equation (DBE)** $\forall x,x'$:
$Q(x)P(x'|x)=Q(x')P(x|x')\implies\pi(x)=\frac{1}{Z}Q(x)$.
Gibbs Sampling: Asympt. correct but slow
1. Init $\mathbf{x}^{(0)}$, fix observed RVs X_B to $\mathbf{x_B}$
2. Repeat: set $\mathbf{x}^{(t)}=\mathbf{x}^{(t-1)}$; select $j\in[1:m]\setminus B$
$x_j^{(t)}\sim P(X_j|\mathbf{x}^{(t)}_{[1:m]\setminus\{j\}})$ (efficient samples)
Random: fulfills DBE, find correct distr.
Determin.: not fulfill DBE, still correct distr.
Expectations via MCMC: Use MCMC sampler (e.g. GS) to get samples $\mathbf{X}^{(1:T)}$. After burn-in time t_0: $\mathbb{E}[f(\mathbf{X})|\mathbf{x}_b]\approx\frac{1}{T-t_0}\sum_{\tau=t_0+1}^T f(\mathbf{X}^{(\tau)})$
Metropolis/Hastings: Generate MC s.t. DBE
1) Proposal $R(X'|X)$, given $X_t=x$, sample $x'\sim R(X'|X=x)$; 2) For $X_t=x$, w.p. $\alpha=\min\{1,\frac{Q(x')R(x|x')}{Q(x)R(x'|x)}\}$: $X_{t+1}=x'$; w.p. $1-\alpha$: $X_{t+1}=x$
Cont RVs: log-concave $p(x)=\frac{1}{Z}exp(-f(x))$, f convex. M/H: $\alpha=\min\{1,\frac{R(x|x')}{R(x'|x)}exp(f(x)-f(x'))$
MALA/LMC: $R(x'|x)=\mathcal{N}(x';x-\tau\nabla f(x);2\tau I)$
→ Use gradient information for convergence
BNN NN weights have distribution
MAP/SGD: $\hat\theta=amin_\theta-\log p(\theta)-\sum_i\log p(y_i|x_i,\theta)$
→ Handles heteroscedastic noise well, fails to predict epistemic uncertainty → use VI
VI(BbB): SGD-opt ELBO via $\nabla_\lambda L(\lambda)$. Find VI approx q_λ. Draw m weights $\theta^{(j)}\sim q_\lambda(\cdot)$. Predict $p(y^*|x^*,x_{1:n},y_{1:n})\approx\frac{1}{m}\sum_j p(y^*|x^*,\theta^{(j)})$
MCMC: produce seq. of weights $\theta^{(1)},..,\theta^{(T)}$ via SGLD, LD, SG-HMC; predict by avg. weights.
Active Learning Get x max. reducing uncertainty
Mutual Info: $I(X;Y)=H(X)-H(X|Y)=I(Y;X)$
Information gain: utility function $f(S), S\subseteq D$,
$F(S):=H(f)-H(f|y_S)=I(f;y_S)=\frac{1}{2}\log|I+\sigma^{-2}K_S|$
Greedy MI optimization: $S_t=\{x_1,..,x_t\}$
$x_{t+1}=\arg\max_{x\in D}F(S_t\cup\{x\})=\arg\max_{x\in D}\sigma^2_{x|S_t}$
Uncertainty sampling: $x_t=\arg\max_{x\in D}\sigma^2_{t-1}(x)$
Heteroscedastic: $\arg\max_{x\in D}\sigma_f^2(x)/\sigma_n^2(x)$
BALD: $x_{t+1}=\arg\max_x I(\theta;y_x|x_{1:t},y_{1:t})=\arg\max_x H(y|x,(x,y)_{1:t})-\mathbb{E}_{\theta\sim p(\cdot|(x,y)_{1:t})}[H(y|x,\theta)]$

Bayesian Optimization Seq. pick $x_1,..,x_T \in D$, get $y_t = f(x_t)+\epsilon_t$, find $\max_x f(x)$ s.t. T small

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

GP-UCB: $x_t = \arg\max_{x\in D} \mu_{t-1}(x)+\beta_t\sigma_{t-1}(x)$ (upper confidence bound ≥ best lower bound) $\mu(x),\sigma(x)$ from GP marginal. β_t EE-tradeoff.

Thm: $f\sim GP$, correct β_t: $\frac{1}{T}R_T = \mathcal{O}^*(\sqrt{\gamma_T/T})$, $\gamma_T = \max_{|S|\le T} I(f;y_S)$ (max. information gain)

EI: choose $x_t = \arg\max_{x\in D} EI(x)$ where $EI(x) = \mathbb{E}[(y^*-y)_+] = \int_{-\infty}^{\infty} max(0,y^*-y)p(y|x)dy$

Thompson sampling: at t, draw from GP post. $\tilde f \sim P(f|x_{1:t},y_{1:t})$, select $x_{t+1}\in\arg\max_{x\in D}\tilde f(x)$

Probab. Planning Control based on prob. model

MDP: A (finite) MDP is defined by States $X=\{1,..,n\}$, Actions $A=\{1,..,m\}$, Transition probabilities $P(x'|x,a)$, Reward function $r(x,a)$ (or $r(x,a,x')$), discount factor $\gamma\in[0,1]$

Planning in MDPs: Policy $\pi: X\to A$ (det.), $\pi: X\to P(A)$ (rand.) induces a MC with transition probabilities $P(X_{t+1}=x'|X_t=x) = P(x'|x,\pi(x))$ (det.) or $\sum_a \pi(a|x)P(x'|x,a)$ (rand.)

Value function: $V^\pi(x) = J(\pi|X_0=x) = \mathbb{E}[\sum_{t=0}^\infty \gamma^t r(X_t,\pi(X_t))|X_0=x] = r(x,\pi(x)) + \gamma\sum_{x'}P(x'|x,\pi(x))V^\pi(x') \Leftrightarrow V^\pi = (I-\gamma T^\pi)^{-1}r^\pi$

$V_i^\pi = V^\pi(i)$, $r_i^\pi = r^\pi(i,\pi(i))$, $T_{i,j}^\pi = P(j|i,\pi(i))$

$V^\pi(x) = \sum_{x'}P(x'|x,\pi(x))[r(x,\pi(x),x')+\gamma V^\pi(x')]$

$V^\pi(x) = Q^\pi(x,\pi(x))$ (deterministic policy π)

$V^\pi(x) = \mathbb{E}_{a'\sim\pi(x)}Q^\pi(x,a')$ (prob. policy $\pi(x)$)

Fixed Point Iter: 1) init V_0^π; 2) for $t=1:T$ do: $V_t^\pi = r^\pi + \gamma T^\pi V_{t-1}^\pi$ (converges)

Greedy policy w.r.t. V: V induces policy $\pi_V(x) = \arg\max_a r(x,a)+\gamma\sum_{x'}P(x'|x,a)V(x')$

Optimal policy: $\pi^* = \arg\max_a Q^*(x,a)$

Bellman Equation: Optimal policy satisfies BE $V^*(x) = \max_{a\in A}\left[r(x,a)+\gamma\sum_{x'\in X}P(x'|x,a)V^*(x')\right] = \max_{a\in A}\mathbb{E}_{x'}[r(x,a)+\gamma V^*(x')] = \max_{a\in A}Q^*(x,a)$

Policy Iteration: 1) Init arbitrary policy π_0 2) Until converged: compute $V^{\pi_t}(x)$; compute greedy policy π_t^G w.r.t. V^{π_t}; set $\pi_{t+1}\leftarrow\pi_t^G$ Stop if $V^{\pi_t}(x) = V^{\pi_{t+1}}(x)$. PI monotonically improves all values $V^{\pi_{t+1}}(x)\ge V^{\pi_t}(x)\forall x$. Finds exact solution in $\mathcal{O}(n^2m/(1-\gamma))$.

Q: $Q_t(x,a) = r(x,a)+\gamma\sum_{x'}P(x'|x,a)V_{t-1}(x')$

Value Iteration: 1) Init $V_0(x) = \max_a r(x,a)$ 2) for $t=1:\infty$: $V_t(x)=\max_a Q_t(x,a)$. Stop if $\|V_t - V_{t-1}\|_\infty \le \epsilon$, then choose greedy π_G w.r.t. V_t. Finds ϵ-opt solution in poly time.

POMDP is a controlled HMM. Can only obtain noisy obsv. Y_t of hidden state X_t. Finite horizon T: exp. #belief states. BUT: most belief states never reached → discretize space by sampling. Use policy gradients with parametric policy.

Belief-state MDP: POMDP as MDP where states ≡ beliefs $P(X_t|y_{1:t})$ in the orig. POMDP. States $\mathcal{B} = \{b:\{1,..,n\}\to[0,1],\sum_{x\in X}b(x)=1\}$, Actions $\mathcal{A}=\{1,..,m\}$, Transitions: $P(Y_{t+1}=y|b_t,a_t) = \sum_{x,x'}b_t(x)P(x'|x,a_t)P(y|x')$; $b_{t+1}(x') = \frac{1}{Z}\sum_x b_t(x)P(X_{t+1}=x'|X_t=x,a_t)P(y_{t+1}|x')$

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

Reinforcement Learning Agent actions change state. State change ~ unknown MDP.

- On-policy: agent has full control (actions)
- Off-policy: no control, only observational data

Model-free RL Directly estimate value function

TD-Learning: (On) Follow π, get (x,a,r,x').

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

Thm: $\alpha_t \vDash RM$ and all (x,a) pairs chosen ∞ often, then $\hat V$ converges to V^π w.p. 1.

Optimistic Q-learning (Off) Estimate $Q^*(x,a)$

1) Init estimate / $Q(x,a) = \frac{R_{max}}{1-\gamma}\prod_{t=1}^{T_{init}}(1-\alpha_t)^{-1}$

2) Pick a (e.g. ϵ_t greedy), get (x,a,r,x'), update: $Q(x,a)\leftarrow(1-\alpha_t)Q(x,a)+\alpha_t(r+\gamma\max_{a'}Q(x',a'))$

Test time: greedy $\pi_G(x)=\arg\max_a Q(x,a)$

Thm: $\alpha_t\vDash RM$, all (x,a) pairs chosen ∞ often, then Q converges to Q^* w.p. 1. Thm(*) holds.

Computation time: $\mathcal{O}(|A|)$, Memory: $\mathcal{O}(|X||A|)$

RL via Function Approx Learn parametric approx. of (action) value function $V(x;\theta), Q(x,a;\theta)$

TD-learning as SGD (On): Tabular TD update rule can be viewed as SGD on loss $l_2(\theta;x,x',r)=\frac12(V(x;\theta)-r-\gamma V(x';\theta_{old}))^2$. Then, $V\leftarrow V-\alpha_t\nabla_{V(x;\theta)}l_2$ is equiv. to TD update.

Function Approx Q-learning (Off) slow

Loss $l_2(\theta;x,a,r,x')=\frac12\delta^2$ where $\delta = Q(x,a;\theta)-r-\gamma\max_{a'}Q(x',a';\theta)$. Alg: Until converged: State x, pick action a, observe r,x'. Update: $\theta\leftarrow\theta-\alpha_t\nabla_\theta l_2 \Leftrightarrow \theta\leftarrow\theta-\alpha_t\delta\nabla_\theta Q(x,a;\theta)$

DQN (Off): Q-learning with NN as func. approx. Use experience replay data D, cloned network to maintain constant NN across episode.

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

Double DQN (Off): Current NN to evaluate action $\arg\max$; prevents maximization bias.

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

$a_t = \arg\max_a Q(x_t,a;\theta)$ intractable for $|A|$ large

Policy Gradient Methods Parametric policy π_θ

Maximize $J(\theta)=\mathbb{E}_{\tau\sim\pi_\theta}[r(\tau)]$ ($\tau = x_{0:T},y_{0:T}$, $r(\tau)=\sum_{t=0}^T\gamma^t r(x_t,a_t)$); via ∇_θ (On). Theorem:

$\nabla_\theta J(\theta)=\nabla_\theta\mathbb{E}_{\tau\sim\pi_\theta}r(\tau)=\mathbb{E}_{\tau\sim\pi_\theta}[r(\tau)\nabla_\theta\log\pi_\theta(\tau)]$

MDP: $\pi_\theta(\tau)=p(x_0)\prod_{t=0}^T\pi(a_t|x_t;\theta)p(x_{t+1}|x_t,a_t)$

Thus: $\nabla_\theta J(\theta)=\mathbb{E}_{\tau\sim\pi_\theta}[r(\tau)\sum_{t=0}^T\nabla_\theta\log\pi(a_t|x_t;\theta)]$

Reducing variance via baselines: $\mathbb{E}_{\tau\sim\pi_\theta}[r(\tau)\nabla\log\pi_\theta(\tau)] = \mathbb{E}_{\tau\sim\pi_\theta}[(r(\tau)-b)\nabla\log\pi_\theta(\tau)]$

Rew2Go: $G_t = \sum_{t'=t}^T\gamma^{t'-t}r_{t'}$; $b_t(x_t)=1/T\sum_{t=0}^{T-1}G_t$

$\nabla J_T(\theta)=\mathbb{E}_{\tau\sim\pi_\theta}[\sum_{t=0}^T\gamma^tG_t\nabla_\theta\log\pi(a_t|x_t;\theta)]$

Mean over returns: replace G_t with $(G_t-b_t(x_t))$

REINFORCE (On): Input $\pi(a|x;\theta)$, init θ

Repeat: generate episode $(x_i,a_i,r_i), i=0:T$; for $t=0:T$: set G_t, update θ:

$\theta=\theta+\eta\gamma^tG_t\nabla_\theta\log\pi(A_t|X_t;\theta)$

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

$\forall x,a: A^{\pi^*}(x,a)\le 0$; $\forall\pi,x:\max_a A^\pi(x,a)\ge0$

Actor-Critic (On) Approx both V^π *and* policy π_θ (e.g. 2 NNs). Reinterpret score gradient:

$\nabla J(\theta_\pi) = \mathbb{E}_{\tau\sim\pi_\theta}[\sum_{t=0}^\infty\gamma^tQ(x_t,a_t;\theta_Q)\nabla\log\pi(a_t|x_t;\theta_\pi)]$

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

Allows online updates:

$\theta_\pi\leftarrow\theta_\pi+\eta_tQ(x,a;\theta_Q)\nabla\log\pi(a|x;\theta_\pi)$

$\theta_Q\leftarrow\theta_Q-\eta_t\delta\nabla Q(x,a;\theta_Q)$ (FA Q-learning)

Variance redution: replace with $Q(x,a;\theta_Q)-V(x;\theta_V)$: advantage func. estimate → A2C

Off-policy Actor Critic (off)

Replace $\max_{a'}Q(x',a';\theta^{old})$ in DQN $L(\theta)$ by $\pi(x';\theta_\pi)$, where π should follow the greedy policy to model $\max_{a'}$. This is equivalent to: $\theta_\pi^*\in\arg\max_\theta\mathbb{E}_{x\sim\mu}[Q(x,\pi(x;\theta);\theta_Q)]$, where $\mu(x)>0$ 'explores all states'. If $Q(\cdot;\theta_Q),\pi(\cdot;\theta_\pi)$ diff'able, use backprop to get stoch. gradients.

$\nabla_\theta J(\theta)=\mathbb{E}_{x\sim\mu}[\nabla_\theta Q(x,\pi(x;\theta);\theta_Q)]$

$\nabla_\theta Q(x,\pi(x;\theta)=\nabla_aQ(x,a)|_{a=\pi(x;\theta)}\cdot\nabla_\theta\pi(x;\theta)$

Needs *deterministic* π. Inject additional action noise (e.g. ϵ_t greedy) to ensure exploration.

Deep Deterministic Policy Gradient (DDPG)

1) init θ_Q,θ_π 2) repeat: observe x, execute $a=\pi(x;\theta_\pi)+\epsilon$, observe r,x', store in D. If time to update: for ITER: sample B from D, compute targets $y=r+\gamma Q(x',\pi(x',\theta_\pi^{old}),\theta_Q^{old})$, update

Critic: $\theta_Q\leftarrow\theta_Q-\eta\nabla 1/|B|\sum_B(Q(x,a;\theta_Q)-y)^2$,

Actor: $\theta_\pi\leftarrow\theta_\pi+\eta\nabla1/|B|\sum_BQ(x,\pi(x;\theta_\pi);\theta_Q)$,

Params: $\theta_j^{old}\leftarrow(1-\rho)\theta_j^{old}+\rho\theta_j$ for $j\in\{\pi,Q\}$

Randomized policy DDPG: For Critic: sample $a'\sim\pi(x';\theta_\pi^{old})$ to get unbiased y estimates. For Actor: consider $\nabla_{\theta_\pi}\mathbb{E}_{a\sim\pi(x;\theta_\pi)}Q(x,a;\theta_Q)$

Reparametrization trick: $a=\psi(x;\theta_\pi,\epsilon)$

$\nabla_{\theta_\pi}\mathbb{E}_{a\sim\pi_{\theta_\pi}}Q(x,a;\theta_Q)=\mathbb{E}_\epsilon\nabla_{\theta_\pi}Q(x,\psi(x;\theta_\pi,\epsilon);\theta_Q)$

Model-based RL Learn MDP, optimize π on it MLE estimate from path trajectory τ:

$P(X_{t+1}|X_t,A)\approx\frac{Cnt(X_{t+1},X_t,A)}{Cnt(X_t,A)}$; $r(x,a)\approx 1/N_{x,a}\sum_{t:X_t=x,A_t=a}R_t$

ϵ_t greedy: Tradeoff exploration-exploitation W.p. ϵ_t: rand. action; w.p. $1-\epsilon_t$: best action. If $\epsilon_t\vDash RM \implies$ converge to π^* w.p. 1.

Robbins Monro (RM): $\sum_t\epsilon_t=\infty$, $\sum_t\epsilon_t^2<\infty$

R_{max} Algorithm: Set unknown $r(x,a)$ to R_{max}, $r(x,a)\le R_{max},\forall x,a$, add fairy tale state x^*, set $P(x^*|x,a)=1$, compute π. Repeat: run π while updating $r(x,a)$, $P(x'|x,a)$, then recompute π.

Thm(*): W.p. $1-\delta$, R_{max} will reach ϵ-opt policy in #steps poly in $|X|,|A|,T,1/\epsilon,\log(1-\delta),R_{max}$.

Note: MDP is assumed ergodic.

Problems of Model-based RL: - Memory required: $P(x'|x,a)\approx\mathcal{O}(|X|^2|A|)$, $r(x,a)\approx\mathcal{O}(|X||A|)$

- Computation: repeatedly solve MDP (VI, PI)

Planning (off) (cont. obsv. states)

MPC (known deterministic dynamics)

Assume known model $x_{t+1}=f(x_t,a_t)$, plan over finite horizon H. At each step t, maximize:

$J_H(a_{t:t+H-1}) := \sum_{\tau=t:t+H-1}\gamma^{\tau-t}r_\tau(x_\tau(a_{t:\tau-1}),a_\tau)$

$x_\tau(a_{t:\tau-1}) = f(f(...(f(x_t,a_t),a_{t+1})..))$

then carry out a_t, then replan.

Optimize via gradient based methods (diff. r,f, cont. action) or via random shooting.

Random shooting: Pick rand. samples $a_{t:t+H-1}^{(i)}$ and pick sample $i^*=\arg\max_i J_H(a_{t:t+H-1}^{(i)})$

MPC with Value estimate: $J_H(a_{t:t+H-1}) := \sum_{\tau=t:t+H-1}\gamma^{\tau-t}r_\tau(x_\tau(a_{t:\tau-1}),a_\tau)+\gamma^HV(x_{t+H})$

$H=1$: $J_1(a_t)=Q(x_t,a_t)$; $\pi_G=\arg\max_aJ_1(a)$

MPC (known stochastic dynamics)

$\max_{a_{t:t+H-1}}\mathbb{E}_{x_{t+1:t+H}}[\sum_{\tau=t:t+H-1}\gamma^{\tau-t}r_\tau+\gamma^HV(x_{t+H})|a_{t:t+H-1}]$

Parametrized policy: ($H=0\Leftrightarrow$ DDPG obj.)

$J_H(\theta)=\mathbb{E}_{x_0\sim\mu}[\sum_{\tau=0:H-1}\gamma^\tau r_\tau+\gamma^HQ(x_H,\pi(x_H,\theta))|\theta]$

MPC (unknown dynamics): follow π, learn f,r,Q off-policy from replay buf, replan π.

BUT: point estimates have poor performance, errors compound → use bayesian learning: Model distribution over f (BNN, GP) and use (approximate) inference (exact, VI, MCMC,..).

Greedy exploitation for model-based RL: (*)

1) $D=\{\}$, prior $P(f|\{\})$ 2) repeat: plan new π to maximize $\max_\pi\mathbb{E}_{f\sim P(\cdot|D)}J(\pi,f)$, rollout π, add new data to D, update posterior $P(f|D)$

PETS algorithm: Ensemble of NNs predicting cond. Gaussian transition distr., use MPC.

Thompson Sampling: Like greedy* BUT in 2) sample model $f\sim P(\cdot|D)$ and then $\max_\pi J(\pi,f)$

Use epistemic noise to drive exploration.

Optimistic exploration: Like greedy* BUT in 2) $\max_\pi\max_{f\in M(D)}J(\pi,f)$; with $M(D)$ set of plausible models given D.