L2

1. MLE 设θ固定，完全依赖data来找参数 $m_h+m_t=m$

$\hat{\theta}_{MLE}=\frac{m_h}{m_t+m_h}$ $L(\theta|D)=\theta^{m_h}(1-\theta)^{m_t}$ 整合所有信息 sufficient statistic

e.g. (m_h, m_t)

但 MLE 无法提供置信度

2. Bayesin：小数据时先验重要，把θ当 RV，在有 prior $P(\theta)$ 下求 $P(\theta|D)$

$P(D_{m+1}=H|D)=\int P(D_{m+1}=H,\theta,D)d\theta=\int P(D_{m+1}=H|\theta,D)P(\theta|D)d\theta$ 只θ，与D无关

Bayesian MAP 简化成求最大后验 $\theta^{MAP}=\arg\max P(\theta|D)\propto\theta^{m_h}(1-\theta)^{m_t}P(\theta)$

Beta分布做先验 $P(\theta)=\frac{\Gamma(\alpha_t+\alpha_h)}{\Gamma(\alpha_t)\Gamma(\alpha_h)}\theta^{\alpha_h-1}(1-\theta)^{\alpha_t-1}$ $E[\theta]=\frac{m_h+\alpha_h}{m+\alpha}$

α_h和α_t是虚构，认为在 $\alpha=\alpha_t+\alpha_h$ 次里的结果

$\Rightarrow\theta^{m_h+\alpha_h-1}(1-\theta)^{m_t+\alpha_t-1}$ $P(D_{m+1}=H|D)=\int\theta P(\theta|D)d\theta=\frac{m_h+\alpha_h}{m+\alpha}$

α↑ ⇒ prior越自信

3. Jensen's Inequality 凹：$\lambda f(x)+(1-\lambda)f(y)\le f(\lambda x+(1-\lambda)y)$

对prob $\sum_{i=1}^n p_i f(x_i)\le f(\sum_{i=1}^n p_i x_i)\Rightarrow E[\log(x)]\le\log(E[x])$

4. Entropy $H(x)=-\sum_{x\in X}P(x)\log p(x)=-E[\log(P(x))]$ $0\log 0=0$

① $H(X)\ge 0$ ② $H(X)=0$ iff $\exists x\in X, P(X=x)=1$ ③ $H(X)\le\log|X|$

proof 3 $H(x)=\frac{1}{n}\sum_{i=1}^n\varphi(p_i)\ge\varphi(\frac{1}{n}\sum_{i=1}^n p_i)=\varphi(\frac{1}{n})=\frac{1}{n}\log n$

$H(p)\le\log n$ 用了 $\varphi(x)=x\log x$ $(x>0)$ $\varphi''(x)>0$ 严格凸 $f(E[X])\le E[f(x)]$

5. 条件entropy $H(Y|X=x)=\sum_y P(y|X=x)\log\frac{1}{P(y|x)}$ 把P(X)换成P(Y|X=x)

H(Y|X=x)可能大于H(Y)，因为X=x可能带来不确定性

平均条件熵 $H(Y|X)=\sum_x p(x)H(Y|X=x)$ 已知X后，Y的剩余平均不确定

联合熵 $H(X,Y)=\sum_{x,y}P(x,y)\log\frac{1}{p(x,y)}$

6. Divergance $D_{KL}(P||Q)=\sum_x P(x)\log\frac{p(x)}{q(x)}$

① ≥ 0. $=-\sum_x P(x)\log\frac{q(x)}{p(x)}\ge-\log(\sum_x P(x)\frac{q(x)}{p(x)})=-\log 1=0$

② $p(x)>0, q(x)=0$, $KL\to\infty$

Cross entropy $H(P,Q)=-\sum_x P(x)\log Q(x)=-E[\log Q(x)]$

$KL(P||Q)=H(P,Q)-H(P)\ge 0\Leftrightarrow$ 即 $\sum_x p(x)\log q(x)\le\sum_x p(x)\log p(x)$

令 $f(x)$ 为非负，有 $\sum_x f(x)\log q(x)\le\sum_x f(x)\log p(x)$ $P^*(x)=f(x)/\sum_x f(x)$

$D_{JS}(P||Q)=\frac{1}{2}KL(P||\frac{P+Q}{2})+\frac{1}{2}KL(Q||\frac{P+Q}{2})$ 有界 $[0,\log 2]$

$\log 2$ 为 P,Q 完全不相交，$\frac{P+Q}{2}(x)$ 退化为 $\frac{1}{2}p(x)$ 或 $\frac{1}{2}q(x)$ $\sum_x p(x)\log 2=\log 2$

最小化 $KL(P||Q)=H(P||Q)-H(P)$ 固定 $\approx-\frac{1}{N}\sum\log Q(x_i)=-\frac{1}{N}\log Q(D)$

MC采样. $P(x,y)$ 乘→ D 当→ $Q(y|x)$ 最小化 $KL\Leftrightarrow\max\sum_{i=1}^N\log Q(y_i|x_i)$

7. Mutual info $I(X;Y)=H(X)-H(X|Y)=\sum_x p(x)\log\frac{1}{p(x)}-\sum_{x,y}p(x,y)\log\frac{1}{p(x|y)}$

$=\sum_{x,y}p(x,y)\frac{p(x|y)}{p(x)}=\sum_{x,y}p(x,y)\frac{p(x,y)}{p(x)p(y)}=KL(P(X,Y)||P(X)P(Y))$

$=H(Y)-H(Y|X)$ 对称 $\Rightarrow I(X,Y)\ge 0$, $=0$ iff $X\perp Y$ 因为是KL

8. C target concept; C concept Class. S训练. H #Hy

泛化 $R(h)=\Pr_{x\sim D}[h(x)\ne c(x)]=E_{x\sim D}[1_{h(x)\ne c(x)}]$ 经验 $\hat{R}_S(h)=\frac{1}{m}\sum 1_{h(x)\ne c(x)}$

9. Prob Appr Corr Lea (PAC) 求 $\Pr[R(h)>\varepsilon]\le\delta$

$m\ge\text{poly}(1/\varepsilon, 1/\delta, n, \text{size}(c))$: $\Pr_{S\sim D^m}[R(h)\le\varepsilon]\ge 1-\delta$ A存在，则PLA-learn

9.1 H=c 空间含 target class 有overlap，放大 一个样本没进去小于 $1-\varepsilon/4$

希望 $\Pr(r_i)\le\varepsilon/4$，这样 放完 $\Pr[\cup r_i]\le\sum\Pr[r_i]$ 小strip m个样本

bad case 有些超 $\varepsilon/4$ $\Pr[\cup_{i=1}^4\Pr(r_i)>\varepsilon/4]\le 4(1-\varepsilon/4)^m\le 4e^{-m\varepsilon/4}$

9.2 H≠c, $R(hs)=0$: consistent. 在训数上能做到 $\hat{R}_S(h)=0$

$\Pr_{S\sim D^m}[R(h)\le\varepsilon]\ge 1-\delta$ holds if $m\ge\frac{1}{\varepsilon}(\log(|H|)+\log\frac{1}{\delta})$, $R(h)\le\frac{1}{m}\log(|H|/\delta\log\frac{1}{\delta})$

Proof: $\Pr[\exists h\in H: \hat{R}_S(h)=0\cap R(h)>\varepsilon]\le\sum_{h\in H}[\hat{R}_S(h)=0\wedge R(h)>\varepsilon]$ 一个放成所有

$=\sum\Pr[\hat{R}_S(h)=0|R(h)>\varepsilon]\Pr[R(h)>\varepsilon]\le|H|(1-\varepsilon)^m\le|H|e^{-m\varepsilon}=\delta$

$\Leftrightarrow m\ge\frac{1}{\varepsilon}(\log(|H|)+\log\frac{1}{\delta})\Leftrightarrow R(h)\le\frac{1}{m}[\log(|H|/\delta+\log\frac{1}{\delta}]$ $O(\frac{1}{m})$

只要 hypothesis set 成立都限，ERM就是PAC

9.3 Inconsistent常见 H≠C且 $R(hs)\ne 0$ Hoeff's ineq X_i独立，$\in[a_i,b_i]$

$\Rightarrow\Pr(\sum X_i-E[\sum X_i]\ge\varepsilon)\le e^{-2\varepsilon^2/\sum(b_i-a_i)^2}$ 和 $\Pr(\sum X_i-E[\sum X_i]\le-\varepsilon)\le e^{-2\varepsilon^2/\sum(b_i-a_i)^2}$

在 Hypo=$\{0,1\}$ 下，代入 $E[\sum X_i]=R(h)$，$\sum X_i$ 是 $\hat{R}(h)$ 代入 $a=0, b=1$

$\Pr[|\hat{R}(h)-R(h)|>\varepsilon]\le 2e^{-2m\varepsilon^2}\Leftrightarrow 2e^{-2m\varepsilon^2}\le\delta\Leftrightarrow|\hat{R}_S(h)-R(h)|\le\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$

$\forall h\in H$, $|\hat{R}_S(h)-R(h)|\ge\varepsilon]\le 2|H|e^{-2m\varepsilon^2}$, $R(h)\le\hat{R}_S(h)+O(\sqrt{\frac{\log|H|}{m}})$ $O(\frac{1}{\sqrt{m}})$

10. Infinite Hypothesis Set (之前 $\log|H|\to\infty$, 发散)

引入 a: growth func b. VC dim $\Pi_H: N\to N$ for a hypothesis H, $\forall m\in N$

$\Pi_H(m)=\max_{\{x_1,\dots,x_m\}\subseteq X}|\{(h(x_1),\dots,h(x_m)\}: h\in H\}|$ 最多的用 $h\in H$ 中把 m 个点分成不同类的数

若H为二分类 $\Pr[|R(h)-\hat{R}(h)|\ge\varepsilon]\le 4\Pi_H(2m)e^{-m\varepsilon^2/8}$

$R(h)\le\hat{R}(h)+\sqrt{\frac{2\log\Pi_H(m)}{m}}+\log\frac{4}{2m}$ 复杂+置信度 $A\subseteq B$, 则 $Vcdim(A)<B$

11. VC Dim ①一种label点的方式 ②能 shattering 能被打散 说明H能实现任意Dichotomy

$VCdim(H)=\max\{m:\Pi_H(m)=2^m\}$ $\Pi_H(m)=2^m$ $VCdim(R^n)=n+1$ 超平面.

$VCdim(\sin)=+\infty$

12. Sauer's lemma, 令H with $VCdim(H)=d$, $\forall m\in N$ $m<d$

$\Pi_H(m)\le\sum_{i=0}^d\binom{m}{i}\le\sum_{i=0}^d\binom{m}{i}\binom{m}{d}^{d-i}=\binom{m}{d}\sum_{i=0}^d\binom{m}{i}\binom{d}{m}^i\le\dots\le(\frac{em}{d})^d$

要么 $VCdim(H)=d<+\infty$, and $\Pi_H(m)=O(m^d)$ 要么 $VCdim(H)=+\infty$, $\Pi_H(m)=2^m$

令 $H\to\{-1,+1\}$ $VCdim(H)=d$, $\forall\delta>0$. 至少有 $1-\delta$ 的 prob $|R(h)-\hat{R}(h)|\le\sqrt{\frac{8d\log\frac{2em}{d}+8\log\frac{4}{\delta}}{m}}$

要让 $|R(h)-\hat{R}(h)|\le\varepsilon$ 对 $\forall h\in H$, with prob p 成立，满足 $m=O(d)$. 模型复杂→数据多

有限 $\forall h\in H$ $|\hat{R}_S(h)-R(h)|\le\sqrt{\frac{\log|H|+\log\frac{2}{\delta}}{2m}}$ $1-\delta$ $O(\frac{1}{\sqrt{m}})$, $\forall h\in H$ $R(h)\le\hat{R}_S(h)+O(\sqrt{\frac{\log|H|}{m}})$

无界 H≠C consist: $R(h)\le\frac{1}{m}(\log|H|+\log\frac{1}{\delta})$ inconsist $|\hat{R}_S(h)-R(h)|\le\sqrt{\frac{\log|H|+\log\frac{2}{\delta}}{2m}}$

Lec3 避前向信号 GD BP层间消失爆

1. 初始化 Xavier $w\sim U[-\frac{\sqrt{6}}{\sqrt{n}},\frac{\sqrt{6}}{\sqrt{n}}]$ n: input dim 每层输入方差大致稳定

He $w\sim N[0,\frac{2}{n}]$ 或 $w\sim N(0,\frac{2}{n})$ 通常用于ReLU 负值无，要更大方差

2. Acti ①Sigmoid $\sigma(z)=\frac{1}{1+\exp(-z)}$ $\sigma'(z)=\sigma(z)(1-\sigma(z))$ vanishing

②ReLU $\max(0,x)$ 快训，但易dead：若某neuron对所有样本output为负. $\sigma=0$

③ReLU变种 $g(z,\alpha)=\max\{0,z\}+\alpha\min\{0,z\}$ $\alpha=-1$ absolute value rectification

$\alpha=0.01$ leaky α 可学 parametric

④ tanh零中心，但saturate $g(z)=\frac{\exp(z)-\exp(-z)}{\exp(z)+\exp(-z)}\in[-1,1]$，比sigmoid

梯度大，当 $|z|\to\infty$, saturate $g'(z)=1-z^2$

⑤ GeLU: $g(z)=z\phi(z)$ ϕ 是 Gaussian CDF softplus $\log(1+\exp(z))$

⑥ Mish $g(z)=z\tanh(\log(1+\exp(z)))=z\tanh(\zeta(z))$

⑦ Swish $g(z)=z\sigma(\beta z)$ β fixed/可学

相同特点：平滑 负半可非0，非单调，上方无界，一般比ReLU稳

3. 输出层 prob匀布

3.1 Linear gaussian $z=W^Th+b$ $P(y|x)=N(y|z,I)$ $L(x,y,\theta)=-\log P(y|x,\theta)$

$\alpha\pm\|y-z\|_2^2$ $\frac{\partial L}{\partial z_k}=z_k-y_k$

3.2 sigmoid =分类 $P(y|x)=\text{Bernoulli}(y|\sigma(z))=\sigma(z)^y(1-\sigma(z))^{1-y}$

$L(x,y,\theta)=-[y\log\sigma(z)+(1-y)\log(1-\sigma(z))]$

$\frac{\partial L}{\partial z}=\frac{\partial L}{\partial\sigma(z)}\cdot\frac{\partial\sigma(z)}{\partial z}=(-\frac{y}{\sigma(z)}+\frac{1-y}{1-\sigma(z)})\cdot\sigma(z)(1-\sigma(z))=\sigma(z)-y$

Saturation 只发生在 model 正确且自信情况下 k是正确标签

3.3 Softmax 多分类 $P(y=k|x)=\frac{\exp(z_k)}{\sum_j\exp(z_j)}$ $L=-\log P(y=k|x)$

对正确角 k $\frac{\partial L}{\partial z_k}=-1+\frac{1}{S}\exp(z^k)=P-1+P(y=k|x)$ $=-z_k+\log S$

预测对吗 对不正确角就没有-1-项

$\frac{\partial L}{\partial z_k}=\hat{y}_k-1(y=k)$ 或 $\frac{\partial L}{\partial z}=-1(y=k)+z$

3.4 BP

$z_k=\sum_j u_j w_{kj}$ 只和k行有关 $u_j=g(z_j)$ $z_j=\sum_i u_i w_{ij}$

$\frac{\partial L}{\partial w_{ij}}=\frac{\partial L}{\partial z_k}\frac{\partial z_k}{\partial w_{kj}}=u_j\delta_k$ $\delta_k=\frac{\partial L}{\partial z_k}$

隐藏 $\frac{\partial L}{\partial w_{ji}}=\sum_k\frac{\partial L}{\partial z_k}\frac{\partial z_k}{\partial w_{kji}}=\sum_k\frac{\partial L}{\partial z_k}\frac{\partial z_k}{\partial u_j}\frac{\partial u_j}{\partial z_j}\frac{\partial z_j}{\partial w_{ji}}$ 部防 $w_{ji}\to z_j\to u_j\to z_k$

$=\sum_k\delta_k W_{kj}g'(z_j)u_i=u_i(g'(z_j))\sum_k\delta_k W_{kj}$ 目标单元误差

定义 $\delta_j=g'(z_j)\sum_k\delta_k W_{kj}$ $\frac{\partial L}{\partial w_{ji}}=u_i\delta_j\frac{\partial L}{\partial b_j}=\delta_j$ 输入激活

4. Optimization

4.1 SGD 只看当前 ∇, noise 大，震荡，狭长 vally，收敛慢

4.2 Momentum $g\leftarrow\frac{1}{m}\nabla_\theta\sum_i L(f(x^{(i)};\theta),y^{(i)})$

$v\leftarrow\alpha v-\varepsilon g$ $\theta\leftarrow\theta+v$ 过去平均指减 $\alpha=0.5, 0.9, 0.99$

4.3 Adagrad $r\leftarrow r+g\odot g$ 对稀疏特征友好，历史∇小的

$\Delta\theta\leftarrow-\frac{\varepsilon}{\delta+\sqrt{r}}\odot g$ $\theta\leftarrow\theta+\Delta\theta$ lr大，cons: r单增，学不动

4.3 RMSProp：遗忘久远历史

$r\leftarrow\rho r+(1-\rho)g\odot g$ $\Delta\theta\leftarrow-\frac{\varepsilon}{\sqrt{\delta+r}}\odot g$ $\theta\leftarrow\theta+\Delta\theta$

指数滑动平均，不会无限累积，有 adaptive scaling 但无显示 momentum. noise仍大

4.4 Adam: Momentum + RMSProp

$s\leftarrow\rho_1 s+(1-\rho_1)g$ $r\leftarrow\rho_2 r+(1-\rho_2)g\odot g$ $\hat{s}=\frac{s}{1-\rho_1^t}$ $\hat{r}=\frac{r}{1-\rho_2^t}$

$\Delta\theta\leftarrow-\varepsilon\frac{\hat{s}}{\sqrt{\hat{r}}+\delta}$ $\theta\leftarrow\theta-\varepsilon\frac{\hat{s}}{\sqrt{\hat{r}}+\delta}$

4.5 LR Schedular：固定 1 step decay/expo decay/cosine warm restart

4.6 Dropout 廉价 bagging $m_j\sim\text{Bernoulli}(p)$ $\tilde{u}_j=m_j u_j$

4.7 convex $O(\frac{1}{\sqrt{T}})$ strongly convex $O(\frac{1}{T})$

$z^{(l)}=W^{(l)}a^{(l-1)}$ forward $a^{(l)}=f(z^{(l)})$

backward $\delta^{(l)}=(W^{(l+1)})^T\delta^{(l+1)}\odot f'(z^{(l)})$

$\frac{\partial L}{\partial W^{(l)}}=\delta^{(l)}(a^{((l-1)})^T$ 上一层激活也是 $W^{(l)}$ 的输入

L4 CNN

输入 filter 位置

1. Convolution $(f*g)(t)=\int_{-\infty}^{+\infty} f(\tau)g(t-\tau)d\tau=\int_{-\infty}^{+\infty} f(t-\tau)g(\tau)d\tau$
$(f*g)(n)=\sum_{m=-\infty}^{+\infty} f[m]g[n-m]=\sum_{m=-\infty}^{+\infty} f[n-m]g[m]$ 卷积核翻转

2. 2D 自相关 $Z(i,j)=(I*K)(i,j)=\sum_m\sum_n I(m,n)K(i-m,j-n)=\sum_m\sum_n I[i-m,j-n]K[m,n]$
$Z(i,j)=(I*K)(i,j)=\sum_m\sum_n I(i+m,j+n)K(m,n)$
DL中 kernel权重可学，不翻转无所谓

3. 三大优势 { Sparse connectivity：一个神经元只连 receptive field 里的
parameter sharing：不同空间相同权重
Equivariance，先移再卷 = 先卷再移

4. output feature channel，l channel
$Z^{(i)}_{j,k}=\sum_{l,m,n} I^{(l)}_{j+m-1,k+n-1}K^{(i,l)}_{m,n}+b^{(i)}$ 加stride
$Z^{(i)}_{j,k}=\sum_{l,m,n} I^{(l)}_{(j-1)s+m,(k-1)s+n}K^{(i,l)}_{m,n}+b^{(i)}$ 一个是 $C_{in}F_hF_w+1$
$H_{out}=\lfloor\frac{H+2P-F_h}{S}\rfloor+1$ 两边 张加口　参数量 $C_{in}F_hF_wC_{out}+C_{out}$
FLOPs = $2\times K_h\times K_w\times C_{in}\times H_{out}\times W_{out}$ 或再加 $H_{out}\times W_{out}$ 次bias相加

5. 反传 J(k,b) 设上一层传回梯度 $G^{(i)}_{j,k}=\frac{\partial J(k,b)}{\partial Z^{(i)}_{j,k}}$ i是第i个output channel
对kernel $\frac{\partial J(k,b)}{\partial k^{(i,l)}_{j,k}}=\sum_{m,n}\frac{\partial J}{\partial Z_{mn}}\frac{\partial Z_{mn}}{\partial k^{(i,l)}_{j,k}}=\sum_{m,n}G^{(i)}_{m,n}I^{(l)}_{(m-1)s+j,(n-1)s+k}$ l是输入chanel
对input $\frac{\partial J(k,b)}{\partial I^{(l)}_{p,q}}=\sum_i\sum_{m,n}K^{(i,l)}_{p-(m-1)s,q-(n-1)s}G^{(i)}_{m,n}$ 每次看到的输入值x那个输出位置，求和
对bias $\frac{\partial J(k,b)}{\partial b^{(i)}}=\sum_{m,n}G^{(i)}_{m,n}$ 一个输入像素 $I_{p,q}$ 会被多个滑窗覆盖，凡覆盖的都要把对应kernel的权重乘上游梯度加起来

6. Pooling { max 平滑 / mean / l_2-norm 能量 / prob weighted }　处理可变尺度输入
对平移不敏感，降计算，关注特征而非位置

7. BN $\mu_j=\frac1N\sum x_{ij}$　$\sigma_j^2=\frac1N\sum(x_{ij}-\mu_j)^2$　$\hat x_{ij}=\frac{x_{ij}-\mu_j}{\sigma_j}$　$\tilde y=\gamma\hat x+\beta$
轻微正则，降初始化敏感，稳深层

8. NIN conv → 1×1 conv → 1×1 conv　GAP：$S_c=\frac{1}{HW}\sum\sum A_{c,i,j}$ 作为
9. Dense：所有前层output全拼接进下一层　类c分数，减参，不用MLP，overfit
10. ResNet X_t 多路径叠再拼　对位置robust，可解释性↑
11. SENet 通道注意力．NasNext 自动找

L5 RNN

1. RNN $h^{(t)}=\tanh(b+Wh^{(t-1)}+Ux^{(t)})$　$o^{(t)}=c+Vh^{(t)}$　$\hat y^{(t)}=softmax(o^{(t)})$
Loss $-\sum_{t=1}^{\tau}P(y^{(t)}|x_{1:t})$ 对所有timestamp求和

2. BPTT $\nabla_W L$　$\nabla_V L$　$\nabla_b L$　$\nabla_c L$　$\nabla_{\theta_m} L$　$\hat y_i(t)=\frac{\exp(o_i^{(t)})}{\sum_j\exp o_j^{(t)}}$
$L^{(t)}=-\log\hat y^{(t)}_{y(t)}$　$\frac{\partial L}{\partial o^{(t)}}=\hat y_i^{(t)}-1_{i,y(t)}$ prob与真实one-hot标
$h^{(t)}$ 到总loss 通过 $o^{(t)}\to L^{(t)}$ 或 $\to h^{(t+1)}\to L^{(t+1)}$ 等
定义 $\delta^{(t)}\equiv\nabla_{a^{(t)}}L$　$\frac{\partial h^{(t)}}{\partial a^{(t)}}=diag(1-(h^{(t)})^2)\Rightarrow\delta^{(t)}=diag(1-h^{(t)2})\nabla_{h^{(t)}}L$
$\Rightarrow\delta^{(t)}=diag(1-(h^{(t)})^2)[V^T\nabla_{o^{(t)}}L+W^T\delta^{(t+1)}]$ 且 $\delta^{(T+1)}=0$ 边界
$\Rightarrow\nabla_b L=\sum_{t=1}^T\delta^{(t)}$，$\nabla_W L=\sum_{t=1}^T\delta^{(t)}(h^{(t-1)})^T$　$\nabla_U L=\sum_{t=1}^T\delta^{(t)}(x^{(t)})^T$

3. Teacher forcing：把 $y^{(t)}$ 当输入 λ $\log p(y_{1:2}|x_{1:2})=\log p(y_2|y_1,x_{1:2})+\log p(y_1|x_{1:2})$ 好处时间步间完全decode ⇒ 可以并行

4. 梯消/爆
线RNN $h^{(t)}=Wh^{(t-1)}+Vx^{(t)}+b$　　非线 $h^{(t)}=f(Wh^{(t-1)}+Vx^{(t)}+b)$
$\frac{\partial h^{(t)}}{\partial h^{(k)}}=\prod_{j=k+1}^t\frac{\partial h^{(j)}}{\partial h^{(j-1)}}=(W^T)^{t-k}$　　$\frac{\partial h^{(t)}}{\partial h^{(j)}}=\prod_{j=k+1}^t W^T diag(f'(h^{(j-1)}))$
若 $W^T=Vdiag(\lambda)V^{-1}$　　$\le\|W^T\|\|diag(\cdots)\|$
$|\lambda_i|>1\to\infty$ 爆
$|\lambda_i|<1\to0$ 消
非线激活 { ① bound recurrent dynamics 缓解爆 ② 若f'很小，sigmoid/tanh → 消失

5. Solu { Gradient clipping $\hat g=\nabla_\theta\varepsilon$　$\hat g\leftarrow\begin{cases}\hat g, & \|\hat g\|<r\\ \frac{r}{\|\hat g\|}\hat g, & \|\hat g\|\ge r\end{cases}$ 防止更新过大
Identity init + ReLU　W=I +ReLU

6. LSTM forget $f^{(t)}=\sigma(W^f h^{(t-1)}+U^f x^{(t)}+b^f)$ 对之前 $c^{(t-1)}$ 留多少
input gate $i^{(t)}=\sigma(W^i h^{(t-1)}+U^i x^{(t)}+b^i)$ 更新什么value
output gate $o^{(t)}=\sigma(W^o h^{(t-1)}+U^o x^{(t)}+b^o)$　candidate values
cell state $c^{(t)}=f^{(t)}\circ c^{(t-1)}+i^{(t)}\circ\tanh(W^c h^{(t-1)}+U^c x^{(t)}+b^c)$
final hidden state $h^{(t)}=o^{(t)}\circ\tanh(c^{(t)})$

7. GRU Update $z^{(t)}=\sigma(W^z h^{(t-1)}+U^z x^{(t)}+b^z)$ 留多少旧hidden state
Reset $r^{(t)}=\sigma(W^r h^{(t-1)}+U^r x^{(t)}+b^r)$ 生成候选是否忘过去
New memory content：$\tilde h^{(t)}=\tanh(W(r^{(t)}\circ h^{(t-1)})+Ux^{(t)}+b)$ 候选记忆
Final memory：$h^{(t)}=z^{(t)}\circ h^{(t-1)}+(1-z^{(t)})\circ\tilde h^{(t)}$
让gradient能沿cell state加法路径传播．
GRU参少 简单训快　LSTM大任务强，长程依赖↑

8. 双向 $\hat y=g(Uh_t+c)=g(U[\vec h_t,\overleftarrow h_t]+c)$，可堆多层

9. Attention $P(y_t|y_{1:t-1},x)=g(y_{t-1},s_t,c_t)$　$s_t=f(s_{t-1},y_{t-1},c_t)$
$c_t=\sum_{j=1}^T\alpha_{tj}h_j$　$\alpha_{tj}=\frac{\exp(e_{tj})}{\sum e_{tk}}$　$e_{tj}=a(s_{t-1},h_j)$

10. 难并行，BP through seq，局部信息与全局信息通过bottleneck
11. 单层 $O(nd^2)$ vs $O(n^2d)$　顺序操作 O(n) vs O(1) 最短路 O(n) vs O(1)
12. Mamba $h_t=A(x_t)h_{t-1}+B(x_t)x_t$　$y_t=C(x_t)h_t$
重要token强更memory 不重要不改．inside a structured state-space evolution

Lec 6

代入 $y-h_S=y-E_S[h_S]+E_S[h_S]-h_S$

1. 分解 $\varepsilon=E_S[E_{(x,y)\sim D}[(y-h_S(x))^2]]=E_SE_{(x,y)}[(y-E_S[h_S]+E_S[h_S]-h_S)^2]$
$=E_SE_{(x,y)}[(y-E_S[h_S])^2]+E_SE_{(x,y)}[(E_S[h_S]-h_S)^2]$
$=E_{(x,y)}[(y-E_S[h_S(x)])^2]+E_SE_x[(E_S[h_S(x)]-h_S(x))^2]$
$=Bias^2+Variance$ 交叉项为0，因为 $E_SE_{(x,y)}[(y-E_S[h_S])(E_S[h_S]-h_S)]$

2. Ridge $J(w,w_0)=\frac1m\sum[y_i-(w_0+w^T\phi(x_i))]^2+\lambda\|w\|_2^2$ 固定x,y，前面只常数
Lasso $J(w,w_0)=\frac1m\sum[y_i-(w_0+w^T\phi(x_i))]^2+\lambda\|w\|_1$　⇒ $\tilde J(\theta)=J(\theta)+\alpha\Omega(\theta)$

3. L2正则/weight decay 等价于对权重加Gaussian Prior后做MAP
$\tilde J(w)=J(w)+\frac\alpha2\|w\|^2$　$w\leftarrow w-\eta(\nabla_wJ(w)+\alpha w)=(1-\eta\alpha)w-\eta\nabla_wJ(w)$ 在未正则化二阶近似
$f(x)\approx f(x_0)+(x-x_0)f'(x_0)+\frac12(x-x_0)^2f''(x)\Rightarrow\hat J(w)=J(w^*)+(w-w^*)\nabla_wJ(w^*)+\frac12(w-w^*)^TH(w-w^*)$
$=J(w^*)+\frac12(w-w^*)^TH(w-w^*)$ H只与w*有关，w*确定，再求导 $\nabla_w\hat J(w)=H(w-w^*)=0$
加入L2后，$\alpha\tilde w+H(\tilde w-w^*)=0$ 令得 $\tilde w=(H+\alpha I)^{-1}Hw^*$ 当α→0 $\tilde w\to w^*$　α↑分解 $H=Q\Lambda Q^T$．Q正交
$\tilde w=(Q\Lambda Q^T+\alpha I)^{-1}Q\Lambda Q^Tw^*=Q(\Lambda+\alpha I)^{-1}\Lambda Q^Tw^*$ 用了 $(ABC)^{-1}=C^{-1}A^{-1}B^{-1}$
在第i个特征方向上，权重被缩放为 $\frac{\lambda_i}{\lambda_i+\alpha}$　L²对低曲率方向(动一点对loss影响小，不重要)惩罚强，在高曲率(重要，影响loss)方向影响弱．了 $w^*=(X^TX+\alpha I)^{-1}X^Ty$ 防止 X^TX 奇异

4. L1-norm $\nabla_wJ(w)=\nabla_wJ(w)+\alpha\,sign(w)$ 不可导，用subgradient．对W*二阶近似，设H对角
$\tilde J(w)=J(w^*)+\sum_i\frac12H_{ii}(w_i-w_i^*)^2+\alpha|w_i|$ 对每个dim $w_i=sign(w_i^*)\max\{|w_i^*|-\frac{\alpha}{H_{ii}},0\}$ 且元素>0
soft-thresholding $|w_i^*|\le\frac{\alpha}{H_{ii}}\Rightarrow w_{ii}=0$ 否则 $|w_i|=|w_i^*|-\frac{\alpha}{H_{ii}}$ 向0移动
$\log p(w)=\sum_i\log Laplacian(w_i;0,\frac1\alpha)=-\alpha\|w\|_1+n\log\alpha-n\log2$

5. Norm Penalty 作为 constrained optimization $\min J(\theta)$，s.t. $\Omega(\theta)-k\le0$ 拉格朗日
$L(\theta,\alpha)=J(\theta)+\alpha(\Omega(\theta)-k)$　$\alpha\ge0\Rightarrow\theta^*=\arg\min_\theta\max_{\alpha,\alpha\ge0}L(\theta,\alpha)$

6. Data Aug：random crop：crop→resize　Dropblock 遮挡，用灰色/随机颜色

7. Noise Injection 等效对权重加惩罚．对input加为Data aug
对weight加 $\tilde w=w+\varepsilon_w$　$\varepsilon_w\sim N(0,\eta I)$　$\hat y(x)=f_w(x)$ 变成 $\hat y_{\varepsilon_w}(x)=f_{w+\varepsilon_w}(x)$
$\tilde J_w=E_{p(x,y,\varepsilon_w)}[(\hat y_{\varepsilon_w}(x)-y)^2]=E_{p(x,y,\varepsilon_w)}[\hat y_{\varepsilon_w}^2-2y\hat y_{\varepsilon_w}+y^2]\approx E_{p(x,y,\varepsilon_w)}[(\hat y_w(x)-y)^2+2(\hat y_w(x)-y)\nabla_w\hat y_w(x)^T\varepsilon_w+(\nabla_w\hat y_w(x)^T\varepsilon_w)^2]$
当η很小，对 $\hat y_{w+\varepsilon_w}(x)\approx\hat y_w(x)+\nabla_w\hat y_w(x)^T\varepsilon_w$ 一阶近似
交叉项依赖 ε_w．$E[\varepsilon_w]=0$ 消去，后面平方项变成 $\eta E_{p(x,y)}[\|\nabla_w\hat y(x)\|^2]$　$\tilde J_w\approx J+\eta E_{p(x,y)}[\|\nabla_w\hat y(x)\|^2]$
右面项measure输出对权重敏感度(L2惩罚 $\|W\|^2$，权重稍变，output变化大)
可以看作是Bayesian inference over weights的 stochastic implementation
w应是一个分布 p(W|D)．精确求解很难，近似做法 训练时不断采样 $w+\varepsilon_w$．不让model 建立在一个极精确值上

8. Output Noise．Label Smoothing 不要对one-hot label过度自信 $y_c^{LS}=1-\varepsilon+\frac{\varepsilon}{k}$　$y_j^{LS}=\frac{\varepsilon}{k}$，j≠c 错误类　$L_{LS}=-\sum_{j=1}^k y_j^{LS}\log p_\theta(j|x)$

Lec7 Transformer

Score = softmax$(\frac{QK^T}{\sqrt{d_k}})V$　pos emb $w=\frac{1}{10000^{2i/d_m}}$
$PE_{pos,2i}=\sin(w_i\,pos)$　$PE_{pos,2i+1}=\cos(w_i\,pos)$　$\begin{bmatrix}\sin w_i(pos+k)\\ \cos w_i(pos+k)\end{bmatrix}=\begin{bmatrix}\cos w_ik & \sin w_ik\\ -\sin w_ik & \cos w_ik\end{bmatrix}\begin{bmatrix}\sin w_ipos\\ \cos w_ipos\end{bmatrix}$
BERT $P(x_i|X_{\setminus i})=softmax(WT_i+b)$　$L_{MLM}=-\sum_{i\in M}\log P(x_i^*|X_{corrupted})$
Next Sentence prediction．50%B在A后，50%随机．使用[CLS]输出向量c预测
$P(IsNext|A,B)=softmax(Wc+b)$　NSP Loss $L_{NSP}=-\log P(y^*_{NSP}|c)$
$L_{BERT}=L_{MLM}+L_{NSP}$　classification 用[CLS]，token-level task
每个token的输出 $\hat y_i=softmax(WT_i+b)$　T_i 第i-th token表示
还有segment embedding 区分开两句子　有pos embed

2. ViT，一个patch $R^{1\times(P^2C)}$　$N=HW/P^2$ 额外加可学习token X_{class} 无embedding
CNN有locality，2D neighbourhood structure，translation equivariance
ViT里只有MLP是local且translationally equivalent．self-attention是global的

3. CLIP：Contrastive Language-Image Pretraining 图文都embed进同一个embedding space
让匹配的相似，不匹配的远离　Loss．对一个patch，含N pair I_i,T_i
$L_{CLIP}=-\frac{1}{2N}\sum_{i=1}^N\log\frac{\exp(I_i\cdot T_i)}{\sum_j\exp(I_i\cdot T_j)}-\frac{1}{2N}\sum_{j=1}^N\log\frac{\exp(I_j\cdot T_j)}{\sum_i\exp(I_i\cdot T_j)}$　image对多text　text对多image
zero shot：先create prompt　$Prob_c=\frac{\exp(I\cdot T_c)}{\sum_{c'=1}^C\exp(I\cdot T_{c'})}$　argmax找最匹配

4. encoder-decoder　masked span prediction
$L_{span}=\sum_{t=1}^{T_y}\log P_\theta(y_t|y_{<t},\tilde x)$．$\tilde x$ 是被mask span替换的序列．y是要恢复的序列

5. LLama：LayerNorm→RMSnorm　ReLU→SwiGLU　MHA→GQA
6. CoT $P_\theta(a|x)=\sum_r P_\theta(a,r|x)=\sum_r P_\theta(r|x)P_\theta(a|x,r)$ 引入中间r，size↑benefit明显
7. LoRA $W'=W+\Delta W=W+BA$　$W\in R^{d\times k}$　$B\in R^{d\times r}$　$A\in R^{r\times k}$　A是Gaussian

Lec9 VAE & Diffusion

1. Intro．两类 { explicit．显式建模 GMM / implicit GAN }　生成 → 处理缺失data，对高维数据handle；→ 结合RL生成env

2. Autoencoder 无监督．target 作为输入本身 $x\in R^n\to h\in R^d$　d<<n 用于降维
encoder $h=f_\theta(x)=a(Wx+b)$　decoder $z=g_{\theta'}(h)=a(W'h+b')$
$\theta^*,\theta'^*=\arg\min\frac1n\sum_i L(x^{(i)},g_{\theta'}(f_\theta(x^{(i)})))$　$L(x,z)=\|x-z\|^2$ 或CE
undercomplete 中间维度比输入小　overcomplete 大　tied weight decoder $W'=W^T$
要加regularization：sparsity，denoising，contractive penalty

3. Linear AE与PCA关系．在线性encoder，l_2 重建error下，最优linear autoencoder
学到的子空间等价于PCA生成的子空间．$A=U\Sigma V^T$　$B^*=\arg\min\|A-B\|_F=U_{:,\le k}\Sigma_{\le k,\le k}V^T$
AE平方误差满足 $\min\sum_t\frac12(x_i^{(t)}-\hat z_i^{(t)})^2\ge\min_{W',h(x)}\frac12\|X-W'h(X)\|_F$　rank(B)≤k　$\Sigma=\frac{1}{m-1}B^TB$
最优解 $W'\leftarrow U_{:,\le k}$　$h(x)\leftarrow\Sigma_{\le k,\le k}V^T_{:,\le k}$
正是PCA $x^{(t)}\leftarrow\frac1n(x^{(t)}-\frac1n\sum x^{(t)})$　$X_{centered}=X-\mu_x$　$\hat X=\frac{1}{\sqrt n}X_{centered}$ 协方差

L9 4. AE→VAE

4.1 显式规定latent space的概率分布⇒破碎连续, 空散

训练时decoder只见过encoder出来的, 没见过能任意采样的

4.2 $z\sim p(z)$ $x\sim p_\theta(x|z)$ $p(z)\sim N(0,I)$ Decoder建模 $p_\theta(x|z)=N(\mu_x, 6_x^2)$

希望最大化 $p_\theta(x)=\int p_\theta(x|z)p(z)dz$ 对所有z, intractable, 且$p_\theta(x|z)$不可解

引入encoder network $q_\phi(z|x)$来近似 $p_\theta(z|x)$. 设 $q_\phi(z|x)=N(\mu_z(x), 6_z(x))$

4.3 ELBO

$\log p_\theta(x)=E_{z\sim q_\phi(z|x)}[\log p_\theta(x)]$ (与z无关) $=E_{z\sim q_\phi(z|x)}[\log \frac{p_\theta(x,z)}{p_\theta(z|x)}]$ 贝叶斯

$=E_{z\sim q_\phi(z|x)}[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\frac{q_\phi(z|x)}{p_\theta(z|x)}]$ 上下同乘 $q_\phi(z|x)$

$=E_{z\sim q_\phi(z|x)}[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}]+D_{KL}(q_\phi(z|x)\|p_\theta(z|x))$ 拆开

$\geq E_{z\sim q_\phi(z|x)}[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}]=L(\theta,\phi,x)$ ($D_{KL}\geq 0$) 最大化ELBO⇔最大化 $\log p_\theta(x)$下界

间接上approximate posterior

$=E_{z\sim q_\phi(z|x)}[\log \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}]$

$=E_{z\sim q_\phi(z|x)}[\log \frac{p(z)}{q_\phi(z|x)}]+E_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)]$

$=-D_{KL}(q_\phi(z|x)\|p(z))+E_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)]$

encoder输出不要背离prior　重建损失

$=\frac{1}{2}\sum_{j=1}^{J}(1+\log 6_{zj}^2-\mu_{zj}^2-6_{zj}^2)$ $\mu_{z,j}$是j-th dim 均值 6方差

1项向于令 $\mu_z\to 0$ $6_z^2\to 1$; $z\sim q_\phi(z|x)$有随机节点, 化为

sample $\varepsilon\sim N(0,I)$ $z=\mu(x)+6(x)\varepsilon$

⇒ $L(\theta,\phi,x)\approx\frac{1}{2}\sum_{j=1}^{J}(1+\log 6_{zj}^2-\mu_{zj}^2-6_{zj}^2)+\frac{1}{L}\sum\log p_\theta(x|z^{(k)})$

前一项对latent dim求和, 后一项对采样次数求和, 用MC

4.4 posterior collapse: 训后encoder学到的posterior退化为prior

z几乎不含x信息. encoder虽然存在, 但被忽略. KL变成0, 正好最大

解决 KL annealing. 初期减KL权重, 学会重构后慢慢加. 限制decoder capacity也行

4.5 VQ-VAE: 离散latent的representation. 一个格子选一个代码.

$q(z=k|x)=\begin{cases}1, & k=\arg\min_j\|z_e(x)-e_j\|_2\\ 0, & \text{otherwise}\end{cases}$ encoder输入量化到最近格子

$L=\log p(x|z_q(x))+\|sg[z_e(x)]-e\|_2^2+\beta\|z_e(x)-sg[e]\|_2^2$

sg[·]前传identity, 后传为0　code book 让靠近encoder输出　commitment 令其输出靠近emb

5. Diffusion

5.1 forward $x_0\sim q(x)$ $x_t=\sqrt{1-\beta_t}x_{t-1}+\sqrt{\beta_t}\varepsilon_t$ $\varepsilon_t\sim N(0,I)$

$q(x_t|x_{t-1})=N(x_t;\sqrt{1-\beta_t}x_{t-1},\beta_t I)$ $q(x_{1:T}|x_0)=\prod_{t=1}^{T}q(x_t|x_{t-1})$

$q(x_{0:T})=q(x_0)\prod_{t=1}^{T}q(x_t|x_{t-1})$

定义 $\alpha_t=1-\beta_t$. $\bar\alpha_t=\prod_{i=1}^{t}\alpha_i$　$z\sim N(\mu_x+\mu_y,6_x^2+6_y^2)$　$x_t=\sqrt{\alpha_t\alpha_{t-1}}x_{t-2}+\sqrt{1-\alpha_t\alpha_{t-1}}\varepsilon$

$x_t=\sqrt{\bar\alpha_t}x_0+\sqrt{1-\bar\alpha_t}\varepsilon$ $\varepsilon\sim N(0,I)$

$q(x_t|x_0)=N(x_t;\sqrt{\bar\alpha_t}x_0,(1-\bar\alpha_t)I)$ t↑⇒ $\bar\alpha_t\to 0$ $q(x_T|x_0)\approx N(0,I)$

5.2 Noise Schedule: linear $\beta_t=\beta_1+\frac{t-1}{T-1}(\beta_T-\beta_1)$

quadratic $\beta_t=[\sqrt{\beta_1}+\frac{t-1}{T-1}(\sqrt{\beta_T}-\sqrt{\beta_1})]^2$

cosine $\bar\alpha_t=\frac{f(t)}{f(0)}$ $f(t)=\cos^2(\frac{t/T+s}{1+s}\cdot\frac{\pi}{2})$　sigmoid $\beta_t=\beta_1+\frac{\beta_T-\beta_1}{1+e^{-(\dots)}}$

5.3 反向. 一步足够小, 仍Gaussian $x_T\sim N(0,I)$ $x_{t-1}=\mu_\theta(x_t,t)+6_t z$

$p_\theta(x_{t-1}|x_t)=N(x_{t-1};\mu_\theta(x_t,t),6_t^2 I)$

5.4 为什么noise prediction等价reverse mean

$q(x_{t-1},x_t|x_0)$是高斯分布, 有其均值 $\mu_q(x_t,x_0)=\frac{\sqrt{\bar\alpha_{t-1}}\beta_t}{1-\bar\alpha_t}x_0+\frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})}{1-\bar\alpha_t}x_t$

$\tilde\mu_t(x_t,x_0)=\frac{1}{\sqrt{\alpha_t}}(x_t-\frac{\beta_t}{\sqrt{1-\bar\alpha_t}}\varepsilon)$ 不妨让模型预测 $\mu_\theta(x_t,t)=\frac{1}{\sqrt{\alpha_t}}(x_t-\frac{\beta_t}{\sqrt{1-\bar\alpha_t}}\varepsilon_\theta(x_t,t))$

最小化两高斯KL等价于让μ_θ接近μ. 等价于让$\varepsilon_\theta(x_t,t)\approx\varepsilon$

$D_{KL}\propto\|\tilde\mu_t-\mu_\theta\|^2\Rightarrow E_{x_0,\varepsilon,t}[\|\varepsilon-\varepsilon_\theta(x_t,t)\|^2]$

怎么得到? $q(x_{t-1}|x_t,x_0)=\frac{q(x_t|x_{t-1})q(x_{t-1}|x_0)}{q(x_t|x_0)}$　$q(x_t|x_{t-1})=\frac{q(x_{t-1}|x_t,x_0)q(x_t|x_0)}{q(x_{t-1}|x_0)}$

log完是 $-\frac{1}{2\beta_t}\|x_t-\sqrt{\alpha_t}x_{t-1}\|^2-\frac{1}{2(1-\bar\alpha_{t-1})}\|x_{t-1}-\sqrt{\bar\alpha_{t-1}}x_0\|^2+\frac{1}{2(1-\bar\alpha_t)}\|x_t-\sqrt{\bar\alpha_t}x_0\|^2$

化简, 只关注x_{t-1}的一次项和二次项. 均值为$A^{-1}B$　$-\frac{1}{2}(\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar\alpha_{t-1}})x_{t-1}^2+(\frac{\sqrt{\alpha_t}}{\beta_t}x_t+\frac{\sqrt{\bar\alpha_{t-1}}}{1-\bar\alpha_{t-1}}x_0)x_{t-1}$　$-\frac{1}{2}Ax^2+Bx$型条件方差A^{-1}

Var: $\frac{\beta_t(1-\bar\alpha_{t-1})}{1-\bar\alpha_t}$　$\mu=\frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})}{1-\bar\alpha_t}x_t+\frac{\sqrt{\bar\alpha_{t-1}}\beta_t}{1-\bar\alpha_t}x_0$

将x_0代入 $x_0=\frac{x_t-\sqrt{1-\bar\alpha_t}\varepsilon}{\sqrt{\bar\alpha_t}}$ ⇒ $\frac{1}{\sqrt{\alpha_t}}(x_t-\frac{\beta_t}{\sqrt{1-\bar\alpha_t}}\varepsilon)$

5.5 ELBO 上下同乘$q(x_{1:T}|x_0)$ 用 $\log E[X]\geq E[\log(X)]$　LVLB

$E_{x_0\sim q(x_0)}[-\log p_\theta(x_0)]\leq E_{x_0,x_{1:T}\sim q(x_{0:T})}[-\log\frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}]$

$=[-\log p(x_T)-\sum_{t=1}^{T}\log p_\theta(x_{t-1}|x_t)+\sum_{t=1}^{T}\log q(x_t|x_{t-1})]$ 展开最后一项

$=[-\log p(x_T)-\sum_{t=1}^{T}\log p_\theta(x_{t-1}|x_t)+\log q(x_T|x_0)-\log q(x_1|x_0)+\sum_{t=2}^{T}\log q(x_{t-1}|x_t,x_0)]$

$=D_{KL}(q(x_T|x_0)\|p(x_T))+\sum_{t=2}^{T}E_q[D_{KL}(q(x_{t-1}|x_t,x_0)\|p_\theta(x_{t-1}|x_t))]-E_q[\log p_\theta(x_0|x_1)]$

让final noisy匹配Gaussian prior　reverse transition matching　reconstruction error

5.6 Train. 取图, t, ε, $\nabla_\theta\|\varepsilon-\varepsilon_\theta(\sqrt{\bar\alpha_t}x_0+\sqrt{1-\bar\alpha_t}\varepsilon,t)\|^2$

Sample: $x_T\sim N(0,I)$ 对每一步 $z\sim N(0,I)$ $x_{t-1}=\frac{1}{\sqrt{\alpha_t}}(x_t-\frac{1-\alpha_t}{\sqrt{1-\bar\alpha_t}}\varepsilon_\theta(x_t,t))$

返回x_0. $\varepsilon_\theta(x_t,t)$　t→embedding→fully-connected注入

Loss $L=E_{x_0\sim q(x_0),t\sim U(1,T),\bar\varepsilon_t\sim N(0,I)}\|\bar\varepsilon_t-\varepsilon_\theta(\sqrt{\bar\alpha_t}x_0+\sqrt{1-\bar\alpha_t}\bar\varepsilon_t,t)\|^2$

$\bar\varepsilon_t$ is the compounded noise added to x_0 to obtain x_t

理论 $6_t^2=\tilde\beta_t^2=\frac{1-\bar\alpha_{t-1}}{1-\bar\alpha_t}\beta_t$ 但实践 $6_t^2=\beta_t$. 若$\beta_t>0$ DDPM. $\beta_t=0$ DDIM

Lec 10

1. ODE SDE

ODE: $\frac{dx}{dt}=f(x,t)$ or $dx=f(x,t)dt$ 解析解 $x(t)=x(0)+\int_0^t f(x,\tau)d\tau$

迭代 $x(t+\Delta t)\approx x(t)+f(x(t),t)\Delta t$ 确定性轨迹. 给定x(0)后续状态完全确定

SDE $\frac{dx}{dt}=f(x,t)+6(x,t)w_t$ drift coefficient　diffusion系数　$dx=f(x,t)dt+6(x,t)dW_t$ random 每次运行都

$x(t+\Delta t)\approx x(t)+f(x(t),t)\Delta t+6(x(t),t)\sqrt{\Delta t}N(0,I)$ 不同, 引入随机噪声

2. 离散diffusion→连续时间ODE

2.1 forward diffusion 连续极限 $x_t=\sqrt{1-\beta_t}x_{t-1}+\sqrt{\beta_t}N(0,I)$ 令$\beta_t:=\beta(t)\Delta t$

$x_t\approx x_{t-1}-\frac{\beta(t)\Delta t}{2}x_{t-1}+\sqrt{\beta(t)\Delta t}N(0,I)$ 当Δt很小时, $\sqrt{1-\beta_t}\approx 1-\frac{1}{2}\beta(t)\Delta t$

令Δt→0 得到前向SDE $dx_t=-\frac{1}{2}\beta(t)x_t dt+\sqrt{\beta_t}dW_t$ (drift　diffusion)

把x_t往0拉. 注入noise

2.2 Reverse time SDE and score function　score = $\nabla_x\log q(x)$

前向: $dx_t=-\frac{1}{2}\beta(t)x_t dt+\sqrt{\beta_t}dW_t$

反向: $dx_t=[-\frac{1}{2}\beta(t)x_t-\beta(t)\nabla_{x_t}\log q_t(x_t)]dt+\sqrt{\beta(t)}d\bar W_t$

Proof: $dx_t=f(x_t,t)dt+g(t)dW_t$

$x_{t+\Delta t}=x_t+f(x_t,t)\Delta t+g(t)\sqrt{\Delta t}\varepsilon$, $\varepsilon\sim N(0,I)$

$p(x'|x)=N(x';x+f(x,t)\Delta t,g(t)^2\Delta t I)$

$p(x|x')\propto p(x'|x)q_t(x)$ 贝. 一阶展开

$\log q_t(x)\approx\log q_t(x')+(x-x')^T\nabla\log q_t(x')$

$E[x_t|x_{t+\Delta t}]=x_{t+\Delta t}-f(x_{t+\Delta t},t)\Delta t+g(t)^2\nabla\log q_t(x_{t+\Delta t})\Delta t$　Gaussian completion

$\Rightarrow dx_t=[f(x_t,t)-g(t)^2\nabla_{x_t}\log q_t(x_t)]dt+g(t)d\bar W$　代入$f=-\frac{1}{2}\beta(t)x_t$ $g=\sqrt{\beta(t)}$

在diffusion中, 反向生成本质学每个noise下的score $S_\theta(x_t,t)\approx\nabla_{x_t}\log q_t(x_t)$

3. Score matching　但score不可算, 且$q_0(x_0)$未知.

希望 $\min_\theta E_t E_{x_t}[\|S_\theta(x_t,t)-\nabla_{x_t}\log q_t(x_t)\|_2^2]$

但有 $\nabla_{x_t}\log q_t(x_t)=E_{x_0\sim q(x_0|x_t)}[\nabla_{x_t}\log q_t(x_t|x_0)]$

proof: $\log q(x_0|x_t)=\log q_0(x_0)+\log q_t(x_t|x_0)-\log q_t(x_t)$

$\nabla_{x_t}\log q(x_0|x_t)=\nabla_{x_t}\log q(x_t|x_0)+\nabla_{x_t}\log q_t(x_t)$

$\nabla_{x_t}\log q(x_0|x_t)=\frac{\nabla_{x_t}q(x_0|x_t)}{q(x_0|x_t)}$ log trick

固定x_t取条件期望. $\int\nabla_{x_t}q(x_0|x_t)dx_0=\nabla_{x_t}\int q(x_0|x_t)dx_0=0$

$E_{x_0\sim q(x_0|x_t)}[\nabla_{x_t}\log q(x_0|x_t)]=\nabla_{x_t}\log q_t(x_t)$ x_t固定时与x_0无关

$S_\theta(x_t,t)\approx\nabla_{x_t}\log q_t(x_t)$ 或 $\nabla_{x_t}\log q_t(x_t)=E_{x_0}[\nabla_{x_t}\log q_t(x_t|x_0)]$

代入 $x_t=\gamma_t x_0+6_t\varepsilon$. $\nabla_{x_t}\log q_t(x_t|x_0)=-\nabla_{x_t}\frac{(x_t-\gamma_t x_0)^2}{26_t^2}=-\frac{\varepsilon}{6_t}$

$\Rightarrow S_\theta(x_t,t):=-\frac{\varepsilon_\theta(x_t,t)}{6_t}$ $\Rightarrow\min_\theta E_{t,x_0,\varepsilon}[\frac{1}{6_t^2}\|\varepsilon-\varepsilon_\theta(x_t,t)\|_2^2]$

本质仍在学noise, 差了系数

4. Probability Flow ODE 反向生成SDE在分布意义上等价于一个ODE

$dx_t=-\frac{1}{2}\beta(t)[x_t+\nabla_{x_t}\log q_t(x_t)]dt$ 解这个ODE等价于$q_t(x_t)$

当初始化 $q_T(x_T)\approx N(x_T;0,I)$. 可用ODE solver采样

5. Flow Matching

Flow Model $x_0\sim p_{init}$, ODE: $dX_t=u_t^\theta(X_t)dt$　模拟ODE/SDE从t=0到t=1 return X_1

Diffusion $x_0\sim p_{init}$ SDE $dX_t=u_t^\theta(X_t)dt+6_t dW_t$

5.1 Time dependent vector field

$u:R^d\times[0,1]\to R^d$

$(x,t)\mapsto u_t(x)$ 轨迹 $\dot X_t=u_t(X_t)\Leftrightarrow\frac{dX_t}{dt}=u_t(X_t)$

瞬时速度场, 输入x和t, 输出一个运动方向和速度, 沿场把分布运过去

5.2 conditional probability path

给定point $z\in R^d$, 定义cond prob path $p_t(\cdot|z)$满足 从noise一路到目标z

$p_0(\cdot|z)=p_{init}$　$p_1(\cdot|z)=\delta_z$ 集中在z的Dirac delta distribution

若用Gaussian $p_t(\cdot|z)=N(\alpha_t z,\beta_t^2 I_d)$ 控制 从μ:0→z, 方差从1→0

边界$\alpha_0=0$, $\beta_0=1$ ⇒ $p_0(\cdot|z)=N(0,I_d)=p_{init}$　$p_1(\cdot|z)=N(z,0)=\delta_z$

5.3 conditional vector field 对任意cond prob path都存在等价vector field

s.t. $X_0\sim p_{init}$ $\frac{dX_t}{dt}=u_t(X_t|z)\Rightarrow X_t\sim p_t(\cdot|z)=N(\alpha_t z,\beta_t^2 I_d)$

$u_t(x|z)=(\dot\alpha_t-\frac{\dot\beta_t}{\beta_t}\alpha_t)z+\frac{\dot\beta_t}{\beta_t}x$　$\dot\alpha_t=\frac{d\alpha_t}{dt}$　$\dot\beta_t=\frac{d\beta_t}{dt}$

推导. 若有 $x_t=\alpha_t z+\beta_t\varepsilon$, ε fix. 则 $\frac{dx_t}{dt}=\dot\alpha_t z+\dot\beta_t\varepsilon$. 又因 $\varepsilon=\frac{x_t-\alpha_t z}{\beta_t}$

整理 $u_t(x_t|z)=\frac{dx_t}{dt}=(\dot\alpha_t-\frac{\dot\beta_t}{\beta_t}\alpha_t)z+\frac{\dot\beta_t}{\beta_t}x_t$

5.4 Marginal Prob Path and vector field 生成整个数据分布

令$z\sim p_{data}$. 定义marginal prob path $p_t(x)=\int p_t(x|z)p_{data}(z)dz$

对应的marginal vector field $u_t(x)=\int u_t(x|z)\frac{p_t(x|z)p_{data}(z)}{p_t(x)}dz$

$=E_{z\sim p(z|X_t=x)}[u_t(x|z)]$ 在位置x上, 有多目标数据能解释

中间态, 但p_{data}未知, $u_t(x)$无法显式写出 ⇒ $\hat u_t(x)$来 估计$u_t(x)$

5.5 FM Loss $L_{CFM}=E_{t\sim U(0,1),z\sim p_{data},x\sim p_t(\cdot|z)}[\|\hat u_t(x)-u_t(x|z)\|^2]$

最小化L_{CFM} ⇔ 最小化L_{FM}. 训练用条件, 期望学的$u_t(x)$

有 $u_t(x|z)=(\dot\alpha_t-\frac{\dot\beta_t}{\beta_t}\alpha_t)z+\frac{\dot\beta_t}{\beta_t}x_t$ 代入 $x_t=\alpha_t z+\beta_t\varepsilon$

$=\dot\alpha_t z+\dot\beta_t\varepsilon$

$L_{CFM}=E_{t,z,\varepsilon}[\|\hat u_t(\alpha_t z+\beta_t\varepsilon)-(\dot\alpha_t z+\dot\beta_t\varepsilon)\|^2]$

5.6 最简单的optimal transport path 令$\alpha_t=t$ $\beta_t=1-t$ $\dot\alpha_t=1$ $\dot\beta_t=-1$

$p_t(x|z)=N(tz,(1-t)^2 I_d)\Rightarrow x=tz+(1-t)\varepsilon$, $\varepsilon\sim N(0,I_d)$

$u_t(x|z)=\dot\alpha_t z+\dot\beta_t\varepsilon=z-\varepsilon$　$L_{CFM}=E[\|\hat u_t(tz+(1-t)\varepsilon)-(z-\varepsilon)\|^2]$

Training ① 从数据 $z\sim p_{data}$ ② $t\sim U(0,1)$ ③ $\varepsilon\sim N(0,I_d)$

④ $x=tz+(1-t)\varepsilon$ ⑤ 用network预测 $\hat u_t(x)$ ⑥ MSE监督 $z-\varepsilon$

6. DDIM: 少步采样, 允许确定性采样 $x_{t-1}=\sqrt{\bar\alpha_{t-1}}\hat x_0+\sqrt{1-\bar\alpha_{t-1}-6_t^2}\hat\varepsilon_t+6_t z$

$\hat x_0=\frac{x_t-\sqrt{1-\bar\alpha_t}\varepsilon_\theta(x_t,t)}{\sqrt{\bar\alpha_t}}$　$\hat\varepsilon_t=\frac{x_t-\sqrt{\bar\alpha_t}\hat x_0}{\sqrt{1-\bar\alpha_t}}$　从 $x_t=\sqrt{\bar\alpha_t}x_0+\sqrt{1-\bar\alpha_t}\varepsilon$ 来

7. CG, CFG

7.1 Vanilla. 给定条件y, 直接训 $u_t^\theta(x|y)$ ①选prompt y ②$x_0\sim p_{init}$

③模拟 $dX_t=u_t^\theta(X_t|u)dt$ 从t=0到1. 不够多样, 不对齐

7.2 classifier guidance

L11 RL

1. $\{(s,a,r,s')\}$ r: 即时reward s'下一状态

2. MDP含(S, A, P, r) S有限状态空间, A动作空间 $P(s'|s,a)$

$r(s,a)=\sum_{s'}r(s,a,s')P(s'|s,a)$ expected reward

2.1 Markov property $P(s_{t+1}|s_t,a_t,\dots)=P(s_{t+1}|s_t,a_t)$

2.2 在π下, agent交互形成trajectory $s_0,a_0,\dots\dots$

discounted total reward $R^\pi(s_0)=\sum_{t=0}^{\infty}\gamma^t r_t$ $\gamma\in[0,1)$ $=r_0+\gamma r_1+\dots$

3. Value function and Optimal Policy

期望return $V^\pi(s)=E[R^\pi(s)]$

最优π*满足 $V^{\pi^*}(s)\geq V^\pi(s)$, $\forall s,\forall\pi$

optimal state value function $V^*(s)=V^{\pi^*}(s)$

4. Planning: 知道 $P(s'|s,a)$, $r(s,a)$ ⇒ π*

RL: 不知环境. 收集 $\{(s,a,r,s')\}$ ⇒ π*　可直接求$P(s'|s,a)$, $r(s,a)$

5. Value Iteration　backup optimality

planning分两步 $P(s'|s,a),r(s,a)\Rightarrow V^*$　$V^*\Rightarrow\pi$ Bellman

value iter ①初始化 $V_0(s)$

②重复更新 $V_{k+1}(s)=\max_a\{r(s,a)+\gamma\sum_{s'}P(s'|a,s)V_k(s')\}$

直到 $\max_s|V_{k+1}(s)-V_k(s)|\leq\varepsilon$　选最好动作

5.1 Bellman operator是contraction　$|\max a-\max b|\leq\max|a-b|$

$\max_s|V_{k+1}(s)-V_k(s)|\leq\gamma\max_s|V_k(s)-V_{k-1}(s)|$

当γ<1时误差↓ 最终收敛到V*　optimality equation

当$V_k=V^*$时, 再迭代不变 $V^*(s)=\max_a\{r(s,a)+\gamma\sum_{s'}P(s'|s,a)V^*(s')\}$

5.2 Optimal Q-function $Q^*(s,a)=r(s,a)+\gamma\sum_{s'}P(s'|s,a)V^*(s')$

由Q*得到策略 $\pi^*(s)=\arg\max_a Q^*(s,a)$

$V^*(s)=\max_a Q^*(s,a)$

直接对Q做value iteration

$Q_{k+1}(s,a)=r(s,a)+\gamma\sum_{s'}P(s'|s,a)\max_{a'}Q_k(s',a')$

收敛到 $Q^*(s,a)=r(s,a)+\gamma\sum_{s'}P(s'|s,a)\max_{a'}Q^*(s',a')$

$\pi_k(s)=\arg\max_a Q_k(s,a)$

6. Q-learning: Value iteration需知$P(s'|s,a)$, $r(s,a)$ 但现实不知

目标 $\{(s,a,r,s')\}\Rightarrow Q^*(s,a)\Rightarrow\pi^*(s)=\arg\max_a Q^*(s,a)$

在value iteration上通过多次采样, 近似为

$Q_{k+1}(s,a)\approx r(s,a)+\gamma\frac{1}{m}\sum_{j=1}^{m}\max_{a'}Q_k(s_j',a')$ 去掉$P(s'|s,a)$

若只有一样本则 $y_{TD}=r+\gamma\max_{a'}Q(s,a')$ TD target　unbiased 方差大

用 $Q(s,a)\leftarrow(1-\alpha)Q(s,a)+\alpha[r+\gamma\max_{a'}Q(s',a')]$

$\Rightarrow Q(s,a)\leftarrow Q(s,a)+\alpha[r+\gamma\max_{a'}Q(s',a')-Q(s,a)]$

α为学习率. $r+\gamma\max$ TD target　$r+\gamma\max-Q$ TD error

定义 $\delta=r+\gamma\max_{a'}Q(s',a')-Q(s,a)$

$Q(s,a)\leftarrow Q(s,a)+\alpha\delta$　TD target大于 提高估计

6.1 Q-Learning. ①初始化 $Q(s,a)$ ②each episode, 选初始 s
重复 用 ε-greedy 找 action a, 得 r 与 s'. $Q(s,a) \leftarrow Q(s,a) + \alpha\delta$, $s \leftarrow s'$
直到 s 到 terminal 结束 episode. 若每个 (s,a) pair 更新无限多, 则收敛.
ε 选 random

7. 7.1 on-policy 当前执行. Q-learning off policy 用 max. 实际不一定走这个
SARSA $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma Q(s',a') - Q(s,a)]$ on-policy

8. DQN 状态空间太大 ⇒ $Q(s,a;\theta) \approx Q^*(s,a)$ $L(\theta) = (y - Q(s,a;\theta))^2$
$y = r + \gamma \max_{a'} Q(s',a';\theta)$
$L(\theta) = ([r + \gamma \max_{a'} Q(s',a';\theta)] - Q(s,a;\theta))^2$ $\theta \leftarrow \theta - \alpha\nabla_\theta$
但 target 本身也依赖 θ, ⇒ moving target ⇒ learning unstable

9. Target Network
引入目标网络参数 θ^-. TD target: $y = r + \gamma \max_{a'} Q(s',a';\theta^-)$
每隔 C 步 同步一次 $\theta^- \leftarrow \theta$ 用慢更新提供稳定 target

10. Experience Replay
DQN 更新: ①样本利用低, 只用一次 ②相邻 trasition 高度相关, 梯度方差大 ③更新只来自一个 state, 收敛慢
Replay Buffer: $D = \{(s_i, a_i, s_i', r_i)\}$ 每次交互后存入
训练时随机采样 mini batch $B = \{(\ )\}$
$L(\theta) = \sum_j ([r_j + \gamma \max_{a_j'} Q(s_j', a_j'; \theta^-)] - Q(s_j, a_j; \theta))^2$
$\theta \leftarrow \theta - \alpha\nabla_\theta L(\theta)$ 一个 transition 被多次采样, 方差↓ 打破时间相关性, 稳定性↑

11. DQN 高估 Q-value $y = r(s,a) + \gamma \max_{a'} E_{w^-}[Q(s',a';\theta^-, w^-)]$
$E_{w^-}[y_Q] = r(s,a) + \gamma E_{w^-}[\max_{a'} Q(s',a';\theta^-, w^-)]$
$\geq r(s,a) + \gamma \max_{a'} E_{w^-}[Q(s',a;\theta^-, w^-)] = \bar{y}$
$E[\max(X_1, X_2)] \geq \max(E[X_1], E[X_2])$

12. Double DQN
DQN 用 θ^- 选动作 + 评估 ⇒ overestimation. Double DQN 分离
动作用 online $a^* = \arg\max_{a'} Q(s',a';\theta)$
评估用 target $y = r(s,a) + \gamma Q(s', a^*; \theta^-)$
更新 $\theta \leftarrow \theta - \alpha\nabla_\theta (y - Q(s,a;\theta))^2$

13. Prioritied Experience Replay
TD-error: $\delta_i = [r(s_i,a_i) + \gamma \max_{a'} Q(s_i', a'; \theta)] - Q(s_i, a_i; \theta)$
priority 定义为 $p_i = |\delta_i| + \varepsilon$ rank-based $p_i = \frac{1}{\text{rank}(i)}$
采样 prob $P(i) = \frac{p_i^\alpha}{\sum_k p_k^\alpha}$ α 为 priority 强度. α=0. uniform
TD-error 大的样本更"意外"

14. Value-based RL
路径 → $Q(s,a;\theta) \Rightarrow \pi^*(s) \to \arg\max_a Q(s,a;\theta)$
通常 determinstic. 可以 off policy. 可用 replay buffer. 样本效率高
缺陷: 间接优化 policy. 需精准估 Q, 估 Q 不稳 policy 不稳
policy 直接学 $\pi_\theta(a|s)$ 路径 ⇒ $\pi_\theta(a|s)$
直接优化 policy, 自然处理连续动作和随机 policy
但通常 on-policy, 样本效率低, 梯度估计方差大

15. Policy gradient $\pi_\theta(\tau) = p(s_0) \prod_{t=0}^{T} \pi_\theta(a_t|s_t) p(s_{t+1}|s_t, a_t)$ 不可导, 不受 θ 控
$J(\theta) = E_{\tau\sim\pi_\theta(\tau)}[r(\tau)]$ $r(\tau) = \sum_t \gamma^t r_t$ $\theta \leftarrow \theta + \alpha\nabla_\theta J$

$\nabla_\theta J(\theta) = \nabla_\theta \int r(\tau)\pi_\theta(\tau) d\tau = \int r(\tau)\pi_\theta(\tau)\nabla_\theta \log\pi_\theta(\tau) d\tau$
$= E_{\tau\sim\pi_\theta(\tau)}[\nabla_\theta \log\pi_\theta(\tau) r(\tau)]$ 用了 $\nabla_\theta\pi_\theta(\tau) = \pi_\theta(\tau)\nabla_\theta\log\pi_\theta(\tau)$
$\log\pi_\theta(\tau) = \log p(s_0) + \sum_t \log\pi_\theta(a_t|s_t) + \sum_t \log p(s_{t+1}|s_t,a_t)$
对θ求导 $\nabla_\theta \log p(s_0) = 0$ ∇_θ 对环境求导没用
$\Rightarrow \nabla_\theta \log\pi_\theta(\tau) = \sum_t \nabla_\theta \log\pi_\theta(a_t|s_t)$ 代回去
$\nabla_\theta J(\theta) = E_{\tau\sim\pi_\theta}[\sum_t \nabla_\theta \log\pi_\theta(a_t|s_t) r(\tau)]$ 可用 N 条代替 $\frac{1}{N}\sum^N$

16. REINFORCE Algo
①采样轨迹 $\{\tau_i\}_{i=1}^N$ 用当前 policy ② $\nabla_\theta J(\theta) \approx \frac{1}{N}\sum^N\sum^t \nabla_\theta \log\pi_\theta(a_t^i|s_t^i) r(\tau^i)$
若 $r(\tau^i) > 0$ 加 prob <0 减 prob on policy. 但更新后旧轨迹不再严格适用新 policy. 样本效率低
因为采样贵, N 小, 方差大, 训练不稳
①用 causality 降方差. 原始更新让每个动作 a_t 乘整个轨迹 reward $\sum_{t=0}^{T}\gamma^t r_t$. 但动作 a_t 无法影响过去 ⇒ reward-to-go $\sum_{t'=t}^{T}\gamma^{t'} r_{t'}$
梯度 $\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^N \sum_{t=0}^{T} [\nabla_\theta \log\pi_\theta(a_t^i|s_t^i) \sum_{t'=t}^{T}\gamma^{t'} r_{t'}^i]$ 减少无关随机
②用 baselines $\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^N\sum_{t=0}^T \nabla_\theta \log\pi_\theta(a_t^i|s_t^i)(r(\tau^i) - b)$
$b = \frac{1}{N}\sum_{i=1}^N r(\tau^i)$ 看是否高于均值
$\nabla_\theta J(\theta) = E_{\tau\sim\pi_\theta}[\nabla_\theta \log\pi_\theta(\tau) r(\tau)] \Rightarrow E_{\tau\sim\pi_\theta}[\nabla_\theta \log\pi_\theta(\tau)(r(\tau) - b)]$
$E_{\tau\sim\pi_\theta}[\nabla_\theta \log\pi_\theta(\tau) b] = \int \pi_\theta(\tau)\nabla_\theta \log\pi_\theta(\tau) b\, d\tau = b\int\nabla_\theta\pi_\theta(\tau)d\tau$
$= b\nabla_\theta\int\pi_\theta(\tau)d\tau = b\nabla_\theta(1) = 0$

17. 原始 PG 每次只能小更新. 太大 performance collapse. 太小 too slow
17.1 Actor-Critic $V^\pi(s) = E_{a\sim\pi(a|s)}[Q^\pi(s,a)]$ PG + policy evaluation
Actor 学 policy $\pi_\theta(a|s)$ Critic 学 value function $V^\pi(s)$ 或 $Q^\pi(s,a)$
Advantage: $A^\pi(s_t,a_t) = Q^\pi(s_t,a_t) - V^\pi(s_t)$ 相对平均的误差
>0 提升 prob <0 降 $\nabla_\theta J(\theta) = \frac{1}{N}\sum_i\sum_t[\gamma^t \nabla_\theta \log\pi_\theta(a_t^i|s_t^i) A^\pi(s_t^i, a_t^i)]$
$Q^\pi(s,a) = r(s,a) + \gamma\sum_{s'} V^\pi(s') P(s'|s,a)$ 用一个样本 s_{t+1} 近似
$\approx r(s_t,a_t) + \gamma V^\pi(s_{t+1})$ $\approx r(s_t,a_t) + \gamma V^\pi(s_{t+1}) - V^\pi(s_t) = \delta_t$ TD error
Critic 学习 $\hat{A}^\pi(s,a) = r + \gamma\hat{V}_\phi^\pi(s') - \hat{V}_\phi^\pi(s)$ $\theta \leftarrow \theta + \alpha\nabla_\theta \log\pi_\theta(a|s)\hat{A}^\pi(s,a)$
训练数据 $(s_i, r_i + \gamma\hat{V}_\phi^\pi(s_i'))$ l_2-loss $L(\phi) = (\hat{V}_\phi^\pi(s_i) - [r_i + \gamma\hat{V}_\phi^\pi(s_i')])^2$
target 是 $r_i + \gamma\hat{V}_\phi^\pi(s_i')$
Batch Actor Critic 采一批新的 $\{(s_i, a_i, s_i', r_i)\}$. critic 更新
$(s_i, r_i + \gamma\hat{V}_\phi^\pi(s_i'))$. advantage: $\hat{A}^\pi(s_i,a_i) = r_i + \gamma\hat{V}_\phi^\pi(s_i') - \hat{V}_\phi^\pi(s_i)$
actor 更新 $\theta \leftarrow \theta + \alpha\sum \nabla_\theta \log\pi_\theta(a_i|s_i) \hat{A}_\pi(s_i, a_i)$
17.2 A3C. decorrelate data. 不同 learner 探索环境不同区域. actor, critic 共享 network
17.3 PPO $\nabla_\theta J(\theta) \approx \frac{1}{N}\sum\sum[\gamma^t \nabla_\theta \frac{\pi_\theta(a_t^i|s_t^i)}{\pi_{\theta old}(a_t^i|s_t^i)} A^\pi(s_t^i, a_t^i)]$ 以前是 $\pi_\theta(a_t^i|s_t^i)$
$L_t^{CLIP}(\theta) = \min(r_t^i A^\pi(s_t^i,a_t^i), \text{clip}(r_t, 1-\varepsilon, 1+\varepsilon)A^\pi(s_t^i, a_t^i))$ prob ratio log trick ≈1
$\nabla_\theta J(\theta) = \frac{1}{N}\sum\sum[\gamma^t \nabla_\theta L_t^{CLIP}(\theta)]$
17.4 SAC $\pi^* = \arg\max_\pi E_{\tau\sim\pi}[\sum_{t=0}^{T}\gamma^t(r(s_t,a_t,s_{t+1}) + \alpha H(\pi(\cdot|s_t)))]$ temp 在 s_t 下 entropy
entropy↑ ⇒ 策略随机. 探索充分. 偏好高熵策略. 探索能力强
比 DDPG. A3C 稳定且高效

L12. GAN
1. 显式生成 $\theta^* = \arg\max_\theta \sum \log p_\theta(x)$ 高维难算 implicit $z \sim p_z(x)$ $\tilde{x} = G(z)$
2. Loss $J(D) = -\frac{1}{2}E_{x\sim p_{data}(x)}[\log D(x)] - \frac{1}{2}E_{z\sim p_z(z)}[\log(1 - D(G(z)))]$
$G(z;\theta_G)$ $D(x;\theta_d) \in [0,1]$ 靠近 1 认为真
$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1 - D(G(z)))]$
$J(G) = -J(D)$ 零和 当 D 很强 $D(G(z)) \approx 0$ 有效梯度弱, 难学
固定生成器 G, 求最优判别器 固定 G, 最大化 $V(G,D)$
$V(G,D) = \int_x p_{data}(x)\log D(x)dx + \int_z p_z(z)\log(1 - D(G(z)))dx$ $\tilde{x} = G(z)$
$= \int_x [p_{data}(x)\log D(x) + p_g(x)\log(1 - D(x))]dx$. 对固定 x, 最大化
$a\log y + b\log(1-y)$ 求导 $y = \frac{a}{a+b}$ $D_G^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)}$

全局最优 $p_g = p_{data}$ $C(G) = V(G, D^*) = E_{x\sim p(x)}[\log\frac{p(x)}{p(x)+p_g(x)}] + E_{x\sim p_g(x)}[\log\frac{p_g(x)}{p(x)+p_g(x)}]$ 定义 $M(x) = \frac{p_{data} + p_g}{2}$ $C(G) = -\log 4 + KL(p_d \| \frac{p_d + p_g}{2})$
$+ KL(p_g \| \frac{p_d+p_g}{2}) = -\log 4 + 2JS(p_{data} \| p_g)$ 且当仅当 $p_{data} = p_g$ (≥0)
训练. 交替. 先固定 G, 更新 D $\max_D E_{x\sim p_d}[\log(D(x))] + E_{\tilde{x}\sim p_g}[\log(1 - D(\tilde{x}))]$
固定 D, 训 G $\min_G E_{\tilde{x}\sim p_g}[\log(1 - D(\tilde{x}))]$ $z \sim p_z(z) \to \tilde{x} = G(z) \to D(x), D(\tilde{x})$
→ update D, G
3. 训练问题 3.1 消失 ∇. 判别太强 $D(G(z)) \approx 0$
Solu: Non-saturating GAN: 生成 loss 改成 $J(G) = -\frac{1}{2}E_z[\log D(G(z))]$
变成最大化 $\log D(G(z))$ 以前是减少误判别为 fake. 现变成更误判为真
当 $D(G(z)) \approx 0$ 时 $-\log D(G(z))$ 非常大. 梯度强 $6^{-1}(u) = \log\frac{u}{1-u}$ sig
3.2 Maximum Likelihood Game 生成 loss $J(G) = -\frac{1}{2}E_z \exp(6^{-1}(D(G(z))))$
等价于最小化 $\theta^* = \arg\min D_{KL}(p_{data}(x) \| p_{model}(x;\theta))$
3.3 不收敛. 异步更新不保证收敛. 更新互相抵消. 震荡
3.4 Mode Collapse. 只生成数据的一小部份 ⇒ 样本局部真实, 但缺少多样性
4. WGAN JS($p_{data} \| p_g$) 若两分布几乎不 overlap, ∇→0
$W(P_r, P_g) = \inf_{\gamma\in\Pi(P_r,P_g)} E_{(x,y)\sim\gamma}\|x - y\|$
kantorovich-Rubinstein duality $W(P_r, P_g) = \sup_{\|f\|_L \le 1} E_{x\sim P_r}[f(x)] - E_{x\sim P_g}[f(x)]$
5. Unrolled GAN
update of D
$\theta_D = \theta_D + \eta\frac{\partial L(\theta_D, \theta_G)}{\partial\theta_D}$
Update of G
$\theta_D^0 = \theta_D$ $\theta_D^{k+1} = \theta_D^k + \eta^k\frac{\partial L(\theta_D^k, \theta_G)}{\partial\theta_D^k}$
$\theta_G = \theta_G - \eta\frac{\partial L_K(\theta_D^K, \theta_G)}{\partial\theta_G}$

易考: 1. PE 自身不 encode order. 需要 order. PE 相对绝对
2. Attention $O(n^2 d)$ 对 sequence 长度增加难以忍受
3. ViT data hungry, less sample coefficient, 大 paras
但 scailing 强. 更灵活. global modeling. 去掉了 local receptive field. weight sharing prior. hard-coded spatial hierarchy CNN 小数据. limited compute edge device strong local pattern
4. Residual gradient flow. vanishing gradient
identify mapping. learn residule mapping
5. BN 初始化. 稳训. 加速训. 可 lr 个
6. Attention 比 RNN 好. 直接 interact. 无 recurrent paths
$I = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$ $K = \begin{bmatrix} 1 & 0 \\ -1 & 2 \end{bmatrix}$ $\Rightarrow Z = \begin{bmatrix} 7 & 9 \\ 13 & 15 \end{bmatrix}$ $J = \sum_{m,n} Z_{m,n}$
$G = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ $\frac{\partial J}{\partial K} = \begin{bmatrix} 12 & 16 \\ 24 & 28 \end{bmatrix}$ 每一个是 Σ 对应
k_{11} 只访问了 4 位置, 每个位置分别乘那个位置与 k_{11} 贡献出来的梯度求和 (1,1) $k_{11} \to Z_{11}$ (1,2) $k_{11} \to Z_{12}$ G 在 I 上卷积
对 k 的梯度 $I(1,1)G(1,1) + I(1,2)G(1,2)$ 其实就是把 G 当
(2,1) $k_{11} \to Z_{21}$ (2,2) $k_{11} \to Z_{22}$
$I(2,1)G(2,1) + I(2,2)G(2,2)$ 反 kernel
对 I 的梯度 把每个上游梯度 × kernel 撒回去再相加
$\begin{bmatrix} 1 & 0 \\ -1 & 2 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ -1 & 2 \end{bmatrix}$ 叠加 $\Rightarrow \begin{bmatrix} 1 & 1 & 0 \\ 0 & 2 & 2 \\ -1 & 1 & 2 \end{bmatrix}$ 反卷积
$\begin{bmatrix} 1 & 0 \\ -1 & 2 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ -1 & 2 \end{bmatrix}$

sigmoid + binary $\hat{y} = 6(z)$ z 是一个数
$L = -y\log\hat{y} - (1-y)\log(1-\hat{y})$ $\frac{\partial L}{\partial z} = \hat{y} - y$
$\frac{\partial L}{\partial z} = \frac{\partial L}{\partial\hat{y}}\frac{\partial\hat{y}}{\partial z} = -\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}}$ $\frac{\partial\hat{y}}{\partial z} = \hat{y}(1-\hat{y})$
$\Rightarrow \hat{y} - y$
softmax + CE $\hat{y}_i = \frac{e^{z_i}}{\sum_j e^{z_j}}$ $\frac{\partial L}{\partial z_i} = \hat{y}_i - y_i$
$\delta = \hat{y} - y$
$x = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ $W^{(1)} = \begin{bmatrix} 1 & -1 \\ 0 & 2 \end{bmatrix}$ $W^{(2)} = [1\ -2]$
$s^{(1)} = W^{(1)}x$. $h = \tanh(s^{(1)})$ $z = W^{(2)}h$
$L = \frac{1}{2}(z-y)^2, y = 1$ $\delta^{(2)} = \frac{\partial L}{\partial z} = z - y$
$\frac{\partial L}{\partial W^{(2)}} = \delta^{(2)}h^T$ $\delta^{(1)} = (W^{(2)})^T\delta^{(2)} \odot (1 - h^2)$
回传到 $\frac{\partial L}{\partial s}$ 相乘和为
CNN $X = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 0 \\ 2 & 1 & 1 \end{bmatrix}$ $k = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$ $S_{ij} = \sum_{ab} k_{ab}X_{i+a, j+b}$
$A = \text{ReLU}(S)$ $z = [1\ {-1}\ 2\ 0]\,\text{vec}(A)$
$L = \frac{1}{2}(z-y)^2, y = 2$ $S = \begin{bmatrix} 1 & 8 \\ 1 & 0 \end{bmatrix}$ $A = \begin{bmatrix} 1 & 8 \\ 1 & 0 \end{bmatrix}$
$\text{vec}(A) = \begin{bmatrix} 8 \\ 6 \end{bmatrix}$ $z = -5$ $\delta^{(2)} = \frac{\partial L}{\partial z} = z - y = -7$
$\frac{\partial L}{\partial W} = \delta^{(2)}\text{vec}(A)^T$ $\frac{\partial L}{\partial \text{vec}(A)} = W^T\delta^{(2)}$
$\frac{\partial L}{\partial S} = \frac{\partial L}{\partial A} \odot 1_{S>0} = \begin{bmatrix} -7 & 7 \\ -14 & 0 \end{bmatrix}$ $\frac{\partial L}{\partial k} = \begin{bmatrix} 7 & -28 \\ -21 & 0 \end{bmatrix}$
$\frac{\partial L}{\partial I}$ 用 $\frac{\partial L}{\partial S}$ 每个值去反卷积, 加一起
RNN $x_1 = 1$ $x_2 = 2$ $h_t = \tan(a_t)$ $a_t = Ux_t + Wh_{t-1}$ $h_0 = 0$
先算 $\delta^{(2)} = \frac{\partial L}{\partial z} = z - y$ $\frac{\partial L}{\partial V} = \delta^{(2)}h_2$ $\frac{\partial L}{\partial h_2} = W\delta^{(2)}$
$\delta^{(2)} = \frac{\partial L}{\partial a_2} = \frac{\partial L}{\partial h_2}(1 - h_2^2)$ $\frac{\partial L}{\partial h_1} = W\delta_2$
$\delta^{(1)} = \frac{\partial L}{\partial a_1} = \frac{\partial L}{\partial h_1}(1 - h_1^2)$ $\frac{\partial L}{\partial U} = \sum_t \delta^t x_t$
$\frac{\partial L}{\partial W} = \sum_t \delta_t h_{t-1}$
$D_{KL}(P\|Q) = \log\frac{6_2}{6_1} + \frac{6_1 + (\mu_1 - \mu_2)^2}{2 6_2^2} - \frac{1}{2}$ $E[X^TAX] = \mu^TA\mu + tr(A\Sigma)$
$D_{KL}(p\|q) = \frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|} + tr(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^T\Sigma_2^{-1}(\mu_2 - \mu_1) - d$