ESSENTIALS

Matrix/Vector Basics

Multiplication: $C = AB \Leftrightarrow c_{ik} = \sum_{j=1}^{m} a_{ij} b_{jk}$

Orthogonal Matrix: (full rank, square matrix with orthonormal columns) $A^{-1} = A^T$, $AA^T = A^TA = I$ $\det(A) \in \{-1, 1\}$, $\det(A^TA) = 1$. preserves: inner product norm, distance, angle, rank, matrix orthogonality

Inner product: $\langle x, y \rangle = x^T y = \sum_{i=1}^{n} x_i y_i$.

$\langle x \pm y, x \pm y \rangle = \langle x, x \rangle \pm 2\langle x, y \rangle + \langle y, y \rangle$

$\langle x, y+z \rangle = \langle x, z \rangle + \langle x, y \rangle$. $\langle x, y \rangle = \|x\|_2 \|y\|_2 \cos(\theta)$

Outer product: $uv^T \Leftrightarrow (uv^T)_{ij} = u_i v_j$

Norms: $\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{\langle x, x \rangle}$, $\|M\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} m_{ij}^2}$

Probability/ Statistics Basics

$P(x) = \sum_{y \in Y} P(x, y)$ $\quad P(x, y) = P(x \mid y) P(y)$

$\forall y \in Y: \sum_{x \in X} P(x \mid y) = 1$

$P(x \mid y) = \frac{P(x, y)}{P(y)} = \frac{P(y \mid x) \cdot P(x)}{P(y)}$ if $P(y) > 0$

posterior $p(A \mid B) = \frac{\text{prior } P(A) \cdot \text{likelihood } p(B \mid A)}{\text{evidence } p(B)}$

If i.i.d: $P(x_1, \dots, x_n) = \prod_{i=1}^{n} P(x_i)$

If X & Y independent: $P(x \mid y) = P(x) \Leftrightarrow P(y \mid x) = P(y)$

Expectation: $\mathbb{E}[f(x)] = \sum_{x \in X} P(x) f(x) = \int P(x) f(x) \, dx$

Variance: $\mathrm{Var}[x] = \mathbb{E}[(x - \mu_x)^2] = \sum_{x \in X} (x - \mu_x)^2 P(x) = \mathbb{E}[x^2] - \mathbb{E}[x]^2$

Expectation Properties

Linearity: $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$, $\mathbb{E}[aX] = a\mathbb{E}[X]$

For constant: $X = c$ then $\mathbb{E}[X] = c$, so we can take the expectation of unrelated stuff and get the same: $\log p_\theta(x^{(i)}) = \mathbb{E}_{z \sim q_\phi(z \mid x^{(i)})}[\log p_\theta(x^{(i)})]$

Double: For any X: $\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X]$.

Convex Function

$\forall x_1, x_2 \in X, \forall t \in [0,1]$: (also if $\forall x \; f''(x) \geq 0$)

$f(tx_1 + (1-t)x_2) \leq t f(x_1) + (1-t) f(x_2)$

Example: Show that if f is convex, any local optimum is global. Assume there is a local optimum $\hat{x}$ that is not the global optimum x^*, then if we choose t in the ball of the local optimum we know that: $f(t\hat{x} + (1-t)x^*) \geq f(\hat{x})$. Since $f(x^*) < f(\hat{x})$, we have $t \cdot f(\hat{x}) + (1-t) f^*(x) < f(\hat{x})$. So we get $f(t\hat{x} + (1-t)x^*) \geq f(\hat{x}) > t \cdot f(\hat{x}) + (1-t) f^*(x)$ which contradicts the convexity of f.

Jensen Inequality

For convex ϕ: $\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(x)]$ and also

$\forall \alpha_i \geq 0$, $\sum \alpha_i = 1$ and any $x_i > 0$: $\log(\sum_i \alpha_i x_i) \geq \sum_i \alpha_i \log(x_i)$

Matrix calculus (numerator layout)

$$\frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \vdots \\ \frac{\partial y_m}{\partial x} \end{bmatrix} \quad \frac{\partial x}{\partial y} = \begin{bmatrix} \frac{\partial x}{\partial y_1} & \dots & \frac{\partial x}{\partial y_n} \end{bmatrix} \quad \frac{\partial y}{\partial z} = \begin{bmatrix} \frac{\partial y_1}{\partial z_1} & \dots & \frac{\partial y_1}{\partial z_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial z_1} & \dots & \frac{\partial y_m}{\partial z_n} \end{bmatrix}$$

$$\frac{\partial A}{\partial x} = \begin{bmatrix} \frac{\partial a_{11}}{\partial x} & \dots & \frac{\partial a_{1n}}{\partial x} \\ \vdots & \ddots & \vdots \\ \frac{\partial a_{m1}}{\partial x} & \dots & \frac{\partial a_{mn}}{\partial x} \end{bmatrix} \quad \frac{\partial x}{\partial A} = \begin{bmatrix} \frac{\partial x}{\partial a_{11}} & \dots & \frac{\partial x}{\partial a_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial x}{\partial a_{m1}} & \dots & \frac{\partial x}{\partial a_{mn}} \end{bmatrix}$$

1D Derivation Rules

$\frac{\partial (f \cdot g)}{\partial x} = f' \cdot g + f \cdot g'$ $\quad \frac{\partial}{\partial x} \frac{f}{g} = \frac{f' \cdot g - g' \cdot f}{g^2}$ $\quad \frac{\partial f(g(x))}{\partial x} = f'(g(x)) \, g'(x)$

$\frac{\partial x^r}{\partial x} = r \cdot x^{(r-1)}$ $\quad \frac{\partial}{\partial x} \frac{1}{f(x)} = -\frac{f'(x)}{(f(x))^2}$ $\quad (f^g)' = f^g \left(f' \frac{g}{f} + g' \ln(f) \right)$

Multivariate Chain Rule

For function $g: \mathbb{R}^n \to \mathbb{R}^m$, $x \mapsto y$ and $f: \mathbb{R}^m \to \mathbb{R}^k$, $y \mapsto z$ we have: $\frac{\partial z_k}{\partial x_j} = \sum_{i=1}^{m} \frac{\partial z_k}{\partial y_i} \frac{\partial y_i}{\partial x_j}$

$\frac{\partial z_k}{\partial x} = \frac{\partial z_k}{\partial y} \frac{\partial y}{\partial x}$ $\quad \frac{\partial z}{\partial x} = \frac{\partial z}{\partial y} \frac{\partial y}{\partial x}$

$\frac{\partial f(x(t), y(t))}{\partial t} = \frac{\partial f}{\partial x} \frac{\partial x}{\partial t} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial t}$ $\quad x, y, t$ can be scalar, vector or matrix.

Derivatives:

$\frac{\partial}{\partial x}(b^T x) = \frac{\partial}{\partial x}(x^T b) = b$ $\quad \frac{\partial}{\partial x}(x^T x) = 2x$ $\quad \frac{\partial}{\partial x}(x^T A x) = (A^T - A)x$

$\frac{\partial}{\partial x}(b^T A x) = A^T b$ $\quad \frac{\partial}{\partial x}(c^T X b) = cb^T$ $\quad \frac{\partial}{\partial X}(c^T X^T b) = bc^T$

$\frac{\partial}{\partial x}(\|x - b\|_2) = \frac{x - b}{\|x - b\|_2}$ $\quad \frac{\partial}{\partial x}(\|x\|_2^2) = 2x$ $\quad \frac{\partial}{\partial X}(\|X\|_F^2) = 2X$

$\frac{\partial}{\partial x} \log(x) = \frac{1}{x}$ $\quad \frac{\partial}{\partial x} \frac{1}{f(x)} = \frac{-f'(x)}{(f(x))^2}$ $\quad \frac{\partial}{\partial x}|x| = (-1)^{\mathbb{1}_{[x<0]}}$

Kullback-Leibler Divergence

$D_{KL}(P \parallel Q) = -\sum_{x \in X} P(x) \log\left(\frac{Q(x)}{P(x)}\right) = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$

$D_{KL}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$

Is not symmetric: $D_{KL}(p \parallel q) \neq D_{KL}(q \parallel p)$

Entropy: $H(p) = -\sum_i p_i \log p_i$

Cross entropy: $H(p, q) = \mathbb{E}_p(-\log(q)) = H(p) + D_{KL}(p \parallel q)$

Jensen-Shannon Divergence

$D_{JS}(P \parallel Q) = \frac{1}{2} D_{KL}(P \parallel \frac{1}{2}(P+Q)) + \frac{1}{2} D_{KL}(Q \parallel \frac{1}{2}(P+Q))$

NEURAL NETWORK BASICS

Challenges (with images): Viewpoint variations, illumination, deformation, occlusion, background clutter, intra-class variation.

Basic linear classifier: $f(x, W) = Wx + b$

Treat pixels as a high-dimensional space, let the classifier draw hyperplanes through that space that separate the classes. Inner product between image and weights to compute a similarity score.

Multiclass SVM loss: $L^i = \sum_{j \neq y^{(i)}} \max(0, s_j - s_{y^{(i)}} + 1)$

What if we drop $j \neq y^{(i)}$: scaled a bit, can also use mean or square

Regularization: L1 $R(W) = \sum_k \sum_l |W_{k,l}|$, L2 $R(W) = \sum_k \sum_l W_{k,l}^2$

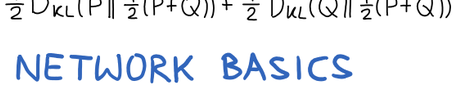
Elastic Net $R(W) = \sum_k \sum_l \beta W_{k,l}^2 + |W_{k,l}|$, MaxNorm, Dropout, ...

Image Features: color histograms (hue), HOG, SIFT features, bag of words (for patches)

Traditional ML: hand-crafted features and kernel, then simple classifier (vs lots of features & learnable kernel)

Perceptron learning: Init $w^{(0)} = 0$, then while $\exists y$ that is misclassified do $\{w^{k+1} = w^k + \eta(y - \hat{y}) x^{i_k}$ where $\hat{y} = ((w^{kT} x^{i_k}) > 0)\}$ If lin. sep. converges in finite time.

Sigmoid function: $\mathrm{sigm}(x) = 1/(1+e^{-x}) = e^x/(e^x+1)$

for $\sigma(w^T x + b)$: 

Logistic regression (MLE): Given $D = \{(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})\}$

Model: $y^{(i)} \sim \mathrm{Bernoulli}(\sigma(w^T x))$.

MLE: $w_{MLE} = \arg\max_w P(D \mid w) = \arg\max_w \prod_{i=1}^{n} P(y^{(i)} \mid x^{(i)}, w)$

$= \arg\max_w \prod_{i=1}^{n} [\underbrace{\sigma(w^T x^{(i)})}_{\pi_i}]^{y^{(i)}} \prod_{i=1}^{n} [\underbrace{1 - \sigma(w^T x^{(i)})}_{1 - \pi_i}]^{1 - y^{(i)}}$

Neg. log. likelihood: $-\log P(D \mid w) = \sum_i y^{(i)} \log \pi_i + (1 - y^{(i)}) \log(1 - \pi_i)$

Softmax function: $\mathrm{softmax}(x_j) = e^{x_j} / \sum_i e^{x_i}$

Loss Functions

Squared Loss: $\frac{1}{2}(y - \hat{y})^2$

Cross entropy: $-y \log(\hat{y}) - (1-y) \log(1-\hat{y})$

Negative log-likelihood: $-\sum_k y_k \log \hat{y}_k$

Why not accuracy: is a step function, so a small change in weights might not change accuracy. Gradient is either zero or undefined.

MLE: Write down prob. dist, decompose into per sample prob., minimize negative log likelihood.

Universal Approximation Theorem

$\exists g(x)$ as NN (with non-linear activation), $g(x) \approx f(x)$ and $|g(x) - f(x)| < \varepsilon \; \forall x \; \forall f \in C^\infty(\mathbb{R})$ (continuous functions) (given enough hidden units, one layer is enough)

Training NN: Init weights to small random values, init biases to 0 or small positive values, update the params with GD. (non-lin. activation make loss non-convex)

TRAINING NEURAL NETWORKS

Regularization: Any technique which aims to reduce generalization error (not training error)

Data Standardization: $X_S = (X - \mu)/\sigma$ where $\mu = \sum_{i=1}^{n} x_i$ and $\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)^2}$. Make sure to use μ and σ for the test set!

Data Augmentation: Generally more data leads to better performance, but acquiring training data is tedious (especially labels). Can generate synthetic data with a model-based approach, but that can lead to discrepancies between the real and synthetic data. Can also expand the data set by transforming existing samples to produce new samples (rotate, scale, noise injection (Gaussian blur, Salt & Pepper noise, channelwise dropout)), need to ensure consistency between transformed data sample and label. For classification introduces invariance $M(T(I)) = M(I)$, for regression introduces equivariance $M(T(I)) = T(M(I))$

Early Stopping: Save parameter where the lowest validation loss occurs then step if there is no improvement for p (patience) epochs. Then retrain model on entire training set for final test error (either for the same number of epochs (easiest & decent performance) or continue training).

As regularizer: Assuming $\|\nabla C(\theta)\| < K$ bounded, learning rate ε, initial parameters θ_0, optimal number of training iterations i^* and param. θ^* learned in i^* steps: $\|\theta_0 - \theta^*\|_2 < K \varepsilon i^*$.

Pseudocode early stopping

```
i, j, i* ← 0 ; v ← ∞ ; θ, θ* ← θ_0   Initialize
while j < p do:   p: patience (how long to wait)
  θ ← train(θ, n) ; i ← i+n   train n steps
  v_c ← validate(θ)
  if v_c < v then
    j ← 0 ; θ* ← θ ; i* ← i ; v ← v_c
  else
    j ← j+1
return θ*, i*   optimal param. w.r.t. validation set
```

Dropout: Let $r_j^l \sim \mathrm{Bernoulli}(p)$ for $p \in [0,1]$

$\tilde{y}^l = r^l \odot y^l$, $z_i^{l+1} = w_i^{l+1} \tilde{y}^l + b_i^{l+1}$, $y_i^{l+1} = f(z_i^{l+1})$ (non-lin. act. func.)

Test time: Would need to evaluate $\sum_r p(r) p(y \mid x, \theta, r)$ (sum over exponentially many terms), but can instead rescale weights: $\tilde{W} = Wp$

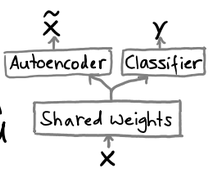
Bagging: Train many models on a bootstrapped dataset, then predict the test set on all models and aggregate (average, majority vote)

Differences to dropout: models are independent (not models sharing weights), train all models until convergence (not train only a small percentage of its models).

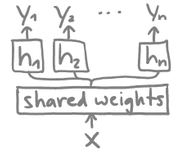
Batch Normalization: **Problem:** in SGD the distribution of x^l constantly changes (internal covariate shift), so we need to adapt constantly. Eliminating this speeds up training as no adaption is necessary.

Solution: standardize input distribution to each layer l: $\bar{x}_k^l = (x_k^l - \mathbb{E}[x_k^l]) / \sqrt{\mathrm{Var}[x_k^l]}$. As this loses some representational power, we perform a linear transform: $\tilde{x}_k^l = \gamma_k^l \bar{x}_k^l + \beta_k^l$ which makes the transform reversible (with learnable parameters γ and β). Note: sometimes $\sqrt{\mathrm{Var}[x_k^l] + \varepsilon}$ for small constant ε (division by 0)

Semi-supervised Learning: Acquiring labeled data is expensive, but unlabeled data usually plentiful. Want to learn from both. Assumes $P(x)$ related to $P(y \mid x)$. Example of this:

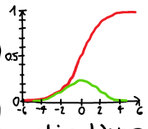


Multitask learning: for datasets with similar samples and related labels. Divide model into shared and task-specific parameters. Assumes that $P(y_1 \mid x)$ related to $P(y_2 \mid x)$, ...



Activation Functions:

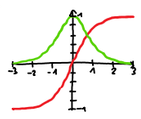
Sigmoid: $\sigma(x) = 1/(1+e^{-x})$, $\sigma'(x) = \sigma(x)(1 - \sigma(x))$



+ interpretation as the firing rate of a neuron. Finite range → stable training
- saturates towards either end, so large activations get small gradients. Not zero-centered, always positive (cannot change "direction").

Tanh: $\tanh(x) = 2\sigma(2x) - 1 = (e^x - e^{-x})/(e^x + e^{-x})$

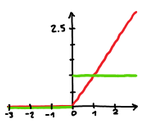
$\tanh'(x) = 1 - \tanh^2$



+ has finite range, zero-centered, has stronger steeper gradients than σ.
- activations also saturate.

Rectified Linear Unit: $\mathrm{ReLU}(x) = \max(0, x)$

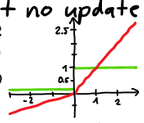
$\mathrm{ReLU}'(x) = \mathbb{1}_{[x>0]} = 1$ if $x > 0$ else 0



+ Inexpensive operation, sparse activations, greatly accelerates convergence.
- can blow up the activation and destabalize training. "dying" ReLU problem: units with negative activations get no update

Leaky ReLU: with small pos. const. α:

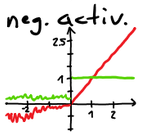
$\mathrm{LReLU}(x) = \begin{cases} \alpha x & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$ $\quad \mathrm{LReLU}'(x) = \begin{cases} \alpha & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$



+ Attempts to solve "dying" ReLU by introducing small negative slope for neg. activ.

Randomized Leaky ReLU: $\alpha \sim U(a, b)$

$f(x) = \begin{cases} \alpha x & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$ $\quad f'(x) = \begin{cases} \alpha & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$



At test time set $\alpha = (a+b)/2$.

Optimization Algorithms:

Gradient Descent: $\theta = \theta - \eta \cdot \nabla_\theta J(\theta)$. Follows the direction of the loss surface downhill.

SGD: $\theta = \theta - \eta \cdot \nabla_\theta J(\theta; x^{(i)}; y^{(i)})$. + prevents redundant calculations, enables to jump to new and potentially better local minima. - has risk of overshooting.

SGD Problem: Gradient big in the direction where we do not need to travel and small in the direction where we want to travel → oscillations, slow convergence

Mini-batch GD: $\theta = \theta - \eta \cdot \nabla_\theta J(\theta; x^{(i:i+n)}; y^{(i:i+n)})$
+ reduces the variance of the parameter updates which can lead to more stable convergence, can make use of highly optimized matrix operations.

Momentum: $v_t = \gamma v_{t-1} + \eta \nabla_\theta J(\theta)$, $\theta = \theta - v_t$. Instead of using the gradient to change the position of a "weight particle", use it to change the velocity. (helps with ravines i.e. areas where the surface curves more steeply in one dimension than the other dimension)

Nesterov accelerated gradient (NAG): $\theta = \theta - v_t$
$v_t = \gamma v_{t-1} + \eta \nabla_\theta J(\theta - \gamma v_{t-1})$. Looks ahead where the momentum would take it, then calculate there

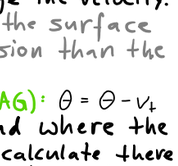
Why adaptive LR: magnitude of gradients can be very different for different layers. Gradients can get very small in the early layers of very deep networks. If a hidden unit has a big fan-in, small changes on many of its incoming weights can cause the learning to overshoot. So we adapt the learning rate to the parameters by performing larger updates for infrequent and smaller update to frequent parameters.

AdaGrad: $\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t,ii} + \varepsilon}} \cdot \nabla_\theta J(\theta_i)$ where $G_t \in \mathbb{R}^{d \times d}$ is a diagonal matrix where each diagonal element i,i is the sum of the squares of the gradients w.r.t. θ_i up to time step t. + Smaller update for larger gradients: avoids overshooting − accumulation of the squared gradients in the denominator keeps growing.

RMS prop/Adadelta: $\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E[g^2]_t + \varepsilon}} \cdot \nabla_\theta J(\theta_i)$, $E[g^2]_t = \gamma E[g^2]_{t-1} + (\gamma - 1) g_t^2$. Attempts to reduce AdaGrad's aggressive, monotonically decreasing learning rate. The sum of squared gradients is recursively defined as a decaying average of all past squared gradients.

Adaptive Moment Estimation (Adam):
$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$ estimate of the first moment
$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$ estimate of the second moment
$\hat{m}_t = m_t / (1-\beta_1^t)$, $\hat{v}_t = v_t / (1-\beta_2^t)$, $\theta_{t+1} = \theta_t - \eta / \sqrt{\hat{v}_t + \varepsilon}\ \hat{m}_t$

Learning rate: If too large, can overshoot, if too low can be slow and get trapped in local minima.



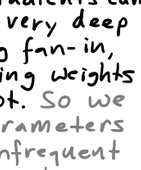
Step decay: decay learning rate by half after certain epochs.
Exponential decay: $\eta_{t+1} = \eta_0 e^{-kt}$
1/t decay: $\eta_{t+1} = \eta_0 / (1 + kt)$

Backpropagation:

$$\delta^{(l-1)} = \frac{\partial C}{\partial z^{(l-1)}} = \frac{\partial C}{\partial z^{(l)}} \frac{\partial z^{(l)}}{\partial z^{(l-1)}} = \sum_{j=1}^{N} \delta_j^{(l)} \frac{\partial z_j^{(l)}}{\partial z_j^{(l-1)}}$$

$$z^{(l)} = \sigma(\theta^{(l)} z^{(l-1)})$$

$$\frac{\partial C}{\partial \theta^{(l)}} = \sum_{j=1}^{N} \frac{\partial C}{\partial z_j^{(l)}} \frac{\partial z_j^{(l)}}{\partial \theta_j^{(l)}} = \sum_{j=1}^{N} \delta_j^{(l)} \frac{\partial z_j^{(l)}}{\partial \theta_j^{(l)}} = \frac{\partial C}{\partial z^{(l)}} \frac{\partial z^{(l)}}{\partial \theta^{(l)}}$$



Receptive fields: particular part of the body surface in which a stimulus will trigger firing in sensory neurons. (everything connected to single input unit form receptive field of that unit)

Receptive Fields Example:

Kernel Size	RF (stride)			
3	1 → (3,1)	→ (5,1)	→ (7,1)	→ (9,1)
3	1 → (3,2)	→ (7,1)	→ (11,1)	→ (15,1)
3	1 → (3,2)	→ (7,2)	→ (15,1)	→ (23,1)

Backprop in MLP

We have $F = F^L \circ F^{L-1} \circ \dots \circ F^1$ and $F(x) = z^L$, $z_i^l = \tanh(w_i^l z^{l-1} + b_i^l)$

We define $\delta^l := \frac{\partial R}{\partial z^l}$ which is:

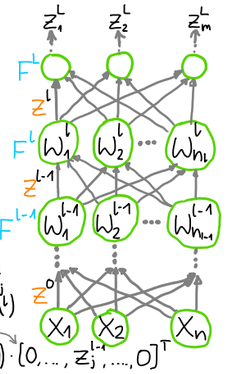
$$\delta^l = \frac{\partial R}{\partial z^l} = \frac{\partial}{\partial z^l}(R \circ F^L \circ \dots \circ F^{l+1}(z^l)) = \delta^{l+1} \cdot \frac{\partial z^{l+1}}{\partial z^l} = \delta^L \cdot \frac{\partial z^L}{\partial z^{L-1}} \cdots \frac{\partial z^{l+1}}{\partial z^l}$$

We can also calculate:

$$\frac{\partial R}{\partial w_{ij}^l} = \frac{\partial}{\partial w_{ij}^l}(R \circ F^L \circ \dots \circ F^{l+1}(z^l)) = \frac{\partial R}{\partial z^l} \frac{\partial z^l}{\partial w_{ij}^l}$$

Let $z^l = \tanh(W^l z^{l-1} + b^l) =: \tanh(a^l)$ (i-th entry)

$$\frac{\partial z^l}{\partial w_{ij}^l} = \frac{\partial \tanh(a^l)}{\partial a^l} \frac{\partial a^l}{\partial w_{ij}^l} = \operatorname{diag}(1 - \tanh^2(a^l)) \cdot [0, \dots, z_j^{l-1}, \dots, 0]^T$$



HMAX Model: simple (S) cells tuned to specific features ($y = \exp(-\frac{1}{2\sigma^2}\sum_{j=1}^{n_{in}}(w_j - x_j)^2)$), complex (c) cells combine output of various S cells to increase invariance and receptive field ($y = \max_{j=1,\dots,c_k} x_j$)

CONVOLUTIONAL NEURAL NETWORKS

Linear Transform: $T(\alpha u + \beta v) = \alpha T(u) + \beta T(v)$
Invariant to f: $T(f(u)) = T(u)$
Equivariant to f: $T(f(u)) = f(T(u))$
↳ any linear, shift-equivariant transform T can be written as a convolution.

Correlation: $I'(i,j) = \sum_{m=-k}^{k} \sum_{n=-k}^{k} K(m,n) I(i+m, j+n)$
Convolution: $I'(i,j) = \sum_{m=-k}^{k} \sum_{n=-k}^{k} K(-m,-n) I(i+m, j+n)$
Called shift-invariant if K does not depend on (i,j)
Same for $K(i,j) = K(-i,-j)$ kernels.
and kernels: matrix-vector multipl. $(I * K) = \begin{bmatrix} k_1 & 0 & \cdots & 0 \\ k_2 & k_1 & & \vdots \\ k_3 & k_2 & \ddots & \\ \vdots & k_3 & & \\ 0 & \vdots & & k_m \end{bmatrix} \begin{bmatrix} I_1 \\ \vdots \\ I_n \end{bmatrix}$

Differentiation: is a linear, shift invariant operation, can write as convolution: $\frac{\partial f}{\partial x} = \lim_{\varepsilon \to 0} \left(\frac{f(x+\varepsilon, y) - f(x,y)}{\varepsilon} \right) \approx \frac{f(x_{n+1}, y) - f(x_n, y)}{\Delta x}$ $\boxed{-1\ 1}$

CNN Backprop: $z^{(l)} = w^{(l)} * z^{(l-1)} + b^{(l)} = \sum_m \sum_n w_{m,n}^{(l)} z_{i-m,j-n}^{(l-1)} + b^{(l)}$

Backward pass w.r.t. inputs

$$\delta_{i,j}^{(l-1)} = \frac{\partial C}{\partial z_{i,j}^{(l-1)}} = \sum_{i'} \sum_{j'} \frac{\partial C}{\partial z_{i',j'}^{(l)}} \frac{\partial z_{i',j'}^{(l)}}{\partial z_{i,j}^{(l-1)}} = \sum_{i'} \sum_{j'} \delta_{i',j'}^{(l)} \frac{\partial}{\partial z_{i,j}^{(l-1)}} \left(\sum_m \sum_n w_{m,n}^{(l)} z_{i'-m,j'-n}^{(l-1)} + b \right)$$

$$= \sum_{i'} \sum_{j'} \delta_{i',j'}^{(l)} \cdot w_{i'-i,j'-j}^{(l)} = \delta^{(l)} * w_{-i,-j} = \delta^{(l)} * \mathrm{ROT}_{180}(w^{(l)})$$

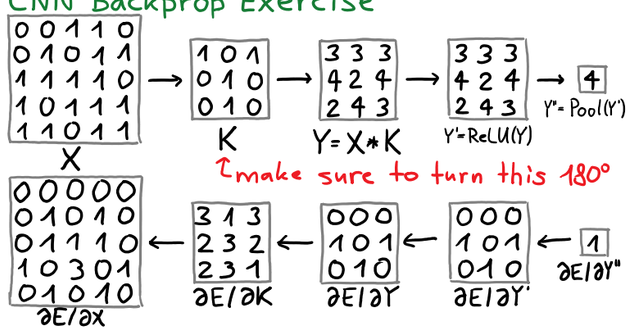
($(w_{00}^{(l)} z_{i-0,j-0}^{(l-1)} + \dots + w_{m,n} z_{i-m,j-n}^{(l-1)} + \dots$ to everything vanishes except where $i=i'-m$, $j=j'-n$. What remains is $w_{m,n}$, but as $i=i'-m \Leftrightarrow i-i'=-m \Leftrightarrow n=j'-n$)

Weight Update:

$$\frac{\partial C}{\partial w_{m,n}^{(l)}} = \sum_i \sum_j \frac{\partial C}{\partial z_{i,j}^{(l)}} \frac{\partial z_{i,j}^{(l)}}{\partial w_{m,n}^{(l)}} = \sum_i \sum_j \delta_{i,j}^{(l)} \frac{\partial}{\partial w_{m,n}^{(l)}} \left(\sum_m \sum_n w_{m,n}^{(l)} z_{i-m,j-n}^{(l-1)} + b^{(l)} \right)$$

$$= \sum_i \sum_j \delta_{i,j}^{(l)} z_{i-m,j-n}^{(l-1)} = \delta^{(l)} * z_{-m,-n}^{(l-1)} = \delta^{(l)} * \mathrm{ROT}_{180}(z^{(l-1)})$$

CNN Backprop Exercise



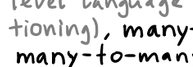
∂E/∂Y'': Just duplicate value where max was
∂E/∂Y': Turn to 1 if greater than 0 else 0
∂E/∂K: Look which elements of X were involved to make that value in Y, then turn that patch 180°, multiply with the corresponding in ∂E/∂Y and add all of these together.
∂E/∂X: Duplicate the 180° turned K at all the locations of ∂E/∂Y. (like upconvolution)

Pooling Layer: makes the representation smaller and more managable. $z^{(l)} = \max\{z_i^{(l-1)}\}$

$$\frac{\partial z^{(l)}}{\partial z^{(l-1)}} = \begin{cases} 1 & \text{if } i = \operatorname{argmax}_i \{z_i^{(l-1)}\} \\ 0 & \text{otherwise.} \end{cases} \qquad \delta^{(l-1)} = \{\delta^{(l)}\}_{i^*} \text{ where } i^* = \operatorname{argmax}_i \{z_i^{(l-1)}\}$$

Deep Networks: many layers: large receptive field despite smaller filters (few parameters)
(A deeper network should always perform at least as good as a shallow network. If not → failure to optimize)
Can be improved with residual connections, reintroducing gradients.

1x1 Convolutions: reduces depth component which leads to fewer parameters, fewer computations. Also increases model capacity.

ResNet: 

Single block: Double convolution with ReLU form block → Stack many on top of each other

Semantic Segmentation: pixel-wise CV task
Using patches or all at once (fully convolutional)
↳ problem: convolutions at full res are expensive, so often uses down and then upsampling.

Unpooling: nearest neighbor (duplicate), bed of nails (only top left, rest 0), max unpooling (put where max was, else 0)

Transpose Convolution: (also de- & up-convolution, fractionally strided conv., backward strided conv.)
Output = copies of filter weighted by input, summed where it overlaps.

Adversarial Defense: Adversarial training (brute force train with lots of adversarial samples to increase robustness), defensive distillation (train with "soft"-class probabilities instead of hard labels. Soft-probabilities stem from 2nd model trained on same task with hard labels. Smoothes network with respect to attack)

RECURRENT NEURAL NETWORKS

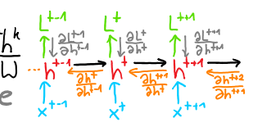
Dynamical System with inputs: $h^t = f(h^{t-1}, x^t; \theta)$
with f: transition function (same for all timesteps)
Instead of $g^t(x^t, \dots, x^1)$, has always same input size

RNN Types: one-to-one (vanilla RNN, character-level language model), one-to-many (image captioning), many-to-one (sentiment classification), many-to-many (delayed) (machine translation), many-to-many (immediate) (video classification on a frame level)

Vanilla RNN: $y^t = W_{hy} h^t$, $h^t = \tanh(W_{hh} h^{t-1} + W_{xh} x^t) = f(h^{t-1}, x^t; W)$
$L^t = \|\hat{Y}^t - Y^t\|^2$, $L = \sum_t L_t$.
Problems with long-term dependencies: (Intuition)
$h^{(t)} = W^T h^{(t-1)} = (W^t)^T h^{(t-1)}$ (note W to the power of t)
if ∃ eigendecomposition of W: $W = Q \Lambda Q^T$ then $h^{(t)} = ((Q \Lambda Q^T)^t)^T h^{(0)}$. If Q is orthogonal: $h^{(t)} = Q^T \Lambda^t Q h^0$
↳ if Λ gets small → vanishes, if Λ gets large → explodes

Backprop through time:

$$\frac{\partial L}{\partial W} = \sum_{t=1}^{\tau} \frac{\partial L^t}{\partial W}, \quad \frac{\partial L^t}{\partial W} = \sum_{k=1}^{t} \frac{\partial L^t}{\partial y^t} \frac{\partial y^t}{\partial h^t} \frac{\partial h^t}{\partial h^k} \frac{\partial^+ h^k}{\partial W}$$



Note: ∂^+ denotes the immediate derivative (keep h^{k-1} fixed)

RNN Backprop Exercise

Given an RNN with formulas $a_t = W h_{t-1} + U x_t + b$
$h_t = \tanh(a_t)$ $\hat{y}_k = \frac{e^{o_{t,k}}}{\sum_j^K e^{o_{t,j}}}$, $L_t = -\sum_k y_{t,k} \log \hat{y}_{t,k}$, $L = \sum_t L_t$
$o_t = V h_t + c$

We want to compute some gradients:

$$\frac{\partial L_t}{\partial o_t} = -\sum_k y_{t,k} \frac{\partial \log \hat{y}_{t,k}}{\partial o_t} = -\sum_k y_{t,k} \frac{\partial \log \hat{y}_{t,k}}{\partial \hat{y}_{t,k}} \frac{\partial \hat{y}_{t,k}}{\partial o_t} = -\sum_k \frac{y_{t,k}}{\hat{y}_{t,k}} \frac{\partial \hat{y}_{t,k}}{\partial o_t}$$

For simplicity we drop the subscript t and compute the partial derivative element-wise:

$$\frac{\partial \hat{y}_k}{\partial o_i} = \frac{\partial}{\partial o_i} \left(\frac{e^{o_k}}{\sum_j^K e^{o_j}} \right) \text{ if } k \neq i \quad \frac{\partial \hat{y}_k}{\partial o_i} = \frac{-e^{o_i} e^{o_k}}{(\sum_j^K e^{o_j})^2} = -\hat{y}_i \cdot \hat{y}_k$$

$$\text{if } k = i \quad \frac{\partial \hat{y}_i}{\partial o_i} = \frac{e^{o_i} \cdot \sum_j^K e^{o_j} - e^{o_i} e^{o_i}}{(\sum_j^K e^{o_j})^2} = \frac{e^{o_i}}{\sum_j^K e^{o_j}} \left(\frac{\sum_j^K e^{o_j}}{\sum_j^K e^{o_j}} - \frac{e^{o_i}}{\sum_j^K e^{o_j}} \right) = \hat{y}_i (1 - \hat{y}_i)$$

We can plug this into equation and rewrite:

$$\frac{\partial L_t}{\partial o_i} = -\frac{y_i}{\hat{y}_i} \cdot (\hat{y}_i (1 - \hat{y}_i)) - \sum_{k \neq i} \frac{y_k}{\hat{y}_k} (-\hat{y}_i \cdot \hat{y}_k) = -y_i + y_i \hat{y}_i + \sum_{k \neq i} y_k \hat{y}_i$$

$$= \hat{y}_i \sum_k y_k - y_i = \hat{y}_i - y_i \quad \text{so} \quad (\nabla_{o_t} L_t)_i = \hat{y}_{t,i} - y_{t,i}$$

$$\left(\frac{\partial L}{\partial o_t} \right)_i = \frac{\partial L}{\partial o_{t,i}} = \frac{\partial L}{\partial L_t} \frac{\partial L_t}{\partial o_{t,i}} = \frac{\partial L_t}{\partial o_{t,i}} \quad \text{since } \partial L / \partial L_t = 1.$$

$$\frac{\partial L_t}{\partial h_t} = \frac{\partial L_t}{\partial o_t} \frac{\partial o_t}{\partial h_t} = \frac{\partial L_t}{\partial o_t} V$$

Note $o = v_1 h_1 + \dots + v_j h_j + \dots + v_m h_m + c$ so the deriv. w.r.t. h_j is v_j.

$$\frac{\partial L}{\partial h_t} = \sum_{t'}^{\tau} \frac{\partial L_{t'}}{\partial h_t} = \frac{\partial L_t}{\partial h_t} + \sum_{t'=t+1}^{\tau} \frac{\partial L_{t'}}{\partial h_t} = \frac{\partial L_t}{\partial o_t} V + \sum_{t'=t+1}^{\tau} \frac{\partial L_{t'}}{\partial h_t}$$

We look at the parts of the sum individually

$$\frac{\partial L_{t'}}{\partial h_t} = \frac{\partial L_{t'}}{\partial h_{t'}} \frac{\partial h_{t'}}{\partial h_{t'-1}} \cdots \frac{\partial h_{t+1}}{\partial h_t} = \frac{\partial L_{t'}}{\partial h_{t'}} \prod_{k=t}^{t'-1} \frac{\partial h_{k+1}}{\partial h_k}$$

We look at a part of the product:

$$\frac{\partial h_{k+1}}{\partial h_k} = \frac{\partial h_{k+1}}{\partial a_{k+1}} \frac{\partial a_{k+1}}{\partial h_k} = \operatorname{diag}(1 - \tanh^2(a_{k+1})) W = \operatorname{diag}(1 - h_{k+1}^2) W$$

The final gradient now is:

$$\frac{\partial L}{\partial h_t} = \frac{\partial L_t}{\partial o_t} V + \sum_{t'=t+1}^{\tau} \left[\frac{\partial L_{t'}}{\partial o_{t'}} V \prod_{k=t}^{t'-1} \operatorname{diag}(1 - h_{k+1}^2) W \right]$$

We can also write it recursively:

$$\frac{\partial L}{\partial h_t} = \frac{\partial L_t}{\partial o_t} V + \operatorname{diag}(1 - h_{t+1}^2) W \frac{\partial L}{\partial h_{t+1}}, \quad \frac{\partial L}{\partial h_\tau} = \frac{\partial L_\tau}{\partial o_\tau} V$$

$$\frac{\partial L}{\partial V} = \sum_t \frac{\partial L_t}{\partial o_t} \frac{\partial o_t}{\partial V}$$ Since h_t does not depend on V, we can treat it as a constant in the definition of o_t.

$$= \sum_t \sum_i \frac{\partial L_t}{\partial o_{t,i}} \frac{\partial o_{t,i}}{\partial V}$$ We look at an individual element V_{kl} in V:

$$\frac{\partial o_{t,i}}{\partial V_{kl}} = \begin{cases} h_l & \text{if } i = k \\ 0 & \text{otherwise} \end{cases} \qquad \frac{\partial o_{t,i}}{\partial V} = \begin{bmatrix} 0 & \cdots & 0 \\ h_{t,1} & \cdots & h_{t,m} \\ 0 & \cdots & 0 \end{bmatrix}$$

Plugging this in and using the result from before:

$$\sum_i \frac{\partial L_t}{\partial o_{t,i}} \frac{\partial o_{t,i}}{\partial V} = \begin{bmatrix} (\hat{y}_{t,1} - y_{t,1}) \cdot h_{t,1} & (\hat{y}_{t,1} - y_{t,1}) \cdot h_{t,2} & \cdots & (\hat{y}_{t,1} - y_{t,1}) h_{t,m} \\ \vdots & & & \vdots \\ (\hat{y}_{t,n} - y_{t,n}) \cdot h_{t,1} & (\hat{y}_{t,n} - y_{t,n}) h_{t,2} & \cdots & (\hat{y}_{t,n} - y_{t,n}) \cdot h_{t,m} \end{bmatrix}$$

$$= \left(\frac{\partial L_t}{\partial o_t} \right)^T \cdot h_t^T \quad \text{so we get:} \quad \frac{\partial L}{\partial V} = \sum_t \left(\frac{\partial L_t}{\partial o_t} \right)^T \cdot h_t^T$$

$$\frac{\partial L_t}{\partial W} = \frac{\partial L_t}{\partial o_t} \frac{\partial o_t}{\partial h_t} \frac{\partial h_t}{\partial W}$$ We now focus on the Last term:

$$\frac{\partial h_t}{\partial W} = \frac{\partial}{\partial W} \left[h_t(W, h_{t-1}(W), h_{t-2}(W), \dots, h_1(W)) \right]$$

$$= \frac{\partial^+ h_t}{\partial W} + \frac{\partial h_t}{\partial h_{t-1}} \frac{\partial^+ h_{t-1}}{\partial W} + \frac{\partial h_t}{\partial h_{t-2}} \frac{\partial^+ h_{t-2}}{\partial W} + \dots + \frac{\partial h_t}{\partial h_1} \frac{\partial^+ h_1}{\partial W}$$

$$= \frac{\partial^+ h_t}{\partial W} + \sum_{k=1}^{t-1} \frac{\partial h_t}{\partial h_k} \frac{\partial^+ h_k}{\partial W} = \frac{\partial h_t}{\partial h_t} \frac{\partial^+ h_t}{\partial W} + \sum_{k=1}^{t-1} \frac{\partial h_t}{\partial h_k} \frac{\partial^+ h_k}{\partial W} = \sum_{k=1}^{t} \frac{\partial h_t}{\partial h_k} \frac{\partial^+ h_k}{\partial W}$$

∂^+ denotes the immediate derivative. All in all:

$$\frac{\partial L_t}{\partial W} = \frac{\partial L_t}{\partial o_t} \frac{\partial o_t}{\partial h_t} \sum_{k=1}^{t} \frac{\partial h_t}{\partial h_k} \frac{\partial^+ h_k}{\partial W} = \sum_{k=1}^{t} \frac{\partial L_t}{\partial o_t} \frac{\partial o_t}{\partial h_t} \frac{\partial h_t}{\partial h_k} \frac{\partial^+ h_k}{\partial W}$$

$\frac{\partial h_t}{\partial h_k}$ for $k < t$: $\frac{\partial h_t}{\partial h_k} = \frac{\partial h_t}{\partial h_{t-1}} \frac{\partial h_{t-1}}{\partial h_{t-2}} \cdots \frac{\partial h_{k+1}}{\partial h_k} = \prod_{t'=k+1}^{t} \frac{\partial h_{t'}}{\partial h_{t'-1}}$

$$= \prod_{t'=k+1}^{t} \frac{\partial h_{t'}}{\partial a_{t'}} \frac{\partial a_{t'}}{\partial h_{t-1}} = \prod_{t'=k+1}^{t} \left[\operatorname{diag}(1 - h_{t'}^2) W \right] = W^{t-k} \prod_{t'=k+1}^{t} \operatorname{diag}(1 - h_{t'}^2)$$

Vanishing / Exploding Gradient:

$\frac{\partial h^t}{\partial h^k} = \prod_{i=k+1}^{t} \frac{\partial h^i}{\partial h^{i-1}} = \prod_{i=k+1}^{t} W_{hh}^T \operatorname{diag}[f'(h^{i-1})]$

We look at a small part of the gradient.

Assume: λ_1 is the largest singular value of W_{hh} and $\|\operatorname{diag}[f'(h^{i-1})]\| < \gamma$ where $\gamma \in \mathbb{R}$ (fine for tanh)

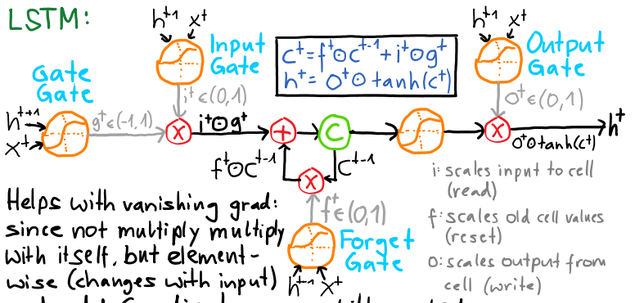
Case 1: It is sufficient $\lambda_1 < 1/\gamma$ for the vanishing gradient problem.

$\forall i, \left\| \frac{\partial h^i}{\partial h^{i-1}} \right\| \leq \|W_{hh}^T\| \cdot \|\operatorname{diag}[f'(h^{i-1})]\| < \frac{1}{\gamma}\gamma = 1$

Let $\eta \in \mathbb{R}$ s.t. $\forall i \ \left\| \frac{\partial h^i}{\partial h^{i-1}} \right\| \leq \eta < 1$. We can see that $\left\| \prod_{i=k}^{t-k} \frac{\partial h^i}{\partial h^{i-1}} \right\| \leq \eta^{(t-k)}$ $\Rightarrow$ this goes to zero for $t \to \infty$.

Case 2: $\lambda_1 > 1/\gamma \Rightarrow$ exploding

LSTM:



Helps with vanishing grad: since not multiply multiply with itself, but element-wise (changes with input) and add. Gradients can still explode.

Gradient clipping:

A way to avoid exploding gradients. Limit the length of the gradient vector

GRU:

Newer, simpler LSTM alternative

$r_t = \sigma(W_{xr} x_t + W_{hr} h_{t-1} + b_r)$, $z_t = \sigma(W_{xz} x_t + W_{hz} h_{t-1} + b_z)$,

$\tilde{h}_t = \tanh(W_{xh} x_t + W_{hh}(r_t \circ h_{t-1}) + b_h)$, $h_t = z_t \circ h_{t-1} + (1 - z_t) \tilde{h}_t$

Image Captioning:

Take some CNN (e.g. AlexNet), drop the fully connected part to label classes (not all FC), append a RNN. Get token probabilities, sample and feed in again (until reaches some END token)

GENERATIVE MODELS

Generative Modelling:

Non-probabilistic (discriminative models) vs probabilistic (generative that are given random vector or discriminative that are also given input (conditionally generative))

(Taxonomy of) Generative models

- Implicit density (highly expressive: universal lack of theory & algorithms)
  - MarkovChain: GSN
  - Direct: GAN
- Explicit density
  - Tractable
    - Belief Nets
    - NADE/MADE
    - PixelRNN/CNN
  - Approximate
    - MarkovChain: Boltzmann Mach.
    - Variational: VAE

Autoencoders:

Finding meaningful DoF that describe the high-dimensional signal with fewer dimensions

encoder f projects input space X into latent space Z

decoder g projects Z back into input space X.

$\hat{\theta}_f, \hat{\theta}_g = \operatorname{argmin}_{\theta_f, \theta_g} \sum_{n=1}^{N} \|x_n - g(f(x_n))\|^2$

If f and g are linear, PCA is optimal solution.

Generative Capabilities: by inducing simple density model over the latent space Z, g can be used to generate new data: $f(x) \sim N(\hat{\mu}, \hat{\sigma} I)$. Problem: gaps in the latent space $\to$ not meaningful samples.

VARIATIONAL AUTOENCODER

Variational Autoencoders:

probabilistic version that allows for sampling from the learned model to generate data with natural variation.

Assume training data D is generated from underlying latent distribution z.

Idea: sample from prior $p_{\theta^*}(z)$ then sample conditional $p_{\theta^*}(x|z^{(i)})$

- x: desired output (e.g. image)
- z: latent factors that control generation process of x (e.g. scale)
- conditional complex (generate image) $\Rightarrow$ param. with neural net.
- choose prior to be a simple distribution (e.g. Gaussian)

VAE Training:

Want to capture this process via estimation of the parameters $p_{\theta^*}(x) = \int_z p_\theta(x|z) p_\theta(z) dz$

- problem: integral over continuous RV. Posterior also intractable
- solution: define additional encoder network $q_\phi(z|x)$ to approx. $p_\theta(x)$ using Bayes' rule.

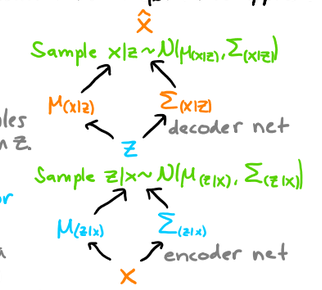
$\mathcal{L}(x^{(i)}, \theta, \phi) = \mathbb{E}_z[\log p_\theta(x^{(i)}|z)]$

(maximize reconstruction likelihood) $\to$ encourages samples of same category to be close in z.

$- D_{KL}(q_\phi(z|x^{(i)}) \| p_\theta(z^{(i)}))$

(make approximate posterior similar to prior)

- needed, otherwise model learns a zero variance (like autoencoder)



- encourage encoder to project latent space representations evenly around the center of the latent space

At run time, only use decoder network, sample z from prior

ELBO Derivation:

start with data log-likelihood:

$\log p_\theta(x^{(i)}) = \mathbb{E}_{z \sim q_\phi(z|x^{(i)})}[\log p_\theta(x^{(i)})]$ ($p_\theta(x^{(i)})$ does not depend on z)

$\mathbb{E}_z\left[\log \frac{p_\theta(x^{(i)}|z) \cdot p_\theta(z)}{p_\theta(z|x^{(i)})} \frac{q_\phi(z|x^{(i)})}{q_\phi(z|x^{(i)})}\right] = \mathbb{E}_z[\log p_\theta(x^{(i)}|z)] - \mathbb{E}_z\left[\log \frac{q_\phi(z|x^{(i)})}{p_\theta(z)}\right] + \mathbb{E}_z\left[\log \frac{q_\phi(z|x^{(i)})}{p_\theta(z|x^{(i)})}\right]$

$= \mathbb{E}_z[\log p_\theta(x^{(i)}|z)] - D_{KL}(q_\phi(z|x^{(i)}) \| p_\theta(z)) + D_{KL}(q_\phi(z|x^{(i)}) \| p_\theta(z|x^{(i)}))$ ≥ 0

Monte Carlo Gradient Estimator

Often encountered: $\nabla_\varphi \mathbb{E}_{p_\varphi(z)}[f(z)]$ with some differentiable function f. We also have a transformation $z = g(\varepsilon, \varphi)$.

$\nabla_\varphi \mathbb{E}_{p_\varphi(z)}[f(z)] = \nabla_\varphi \int p_\varphi(z) f(z) dz = \nabla_\varphi \int p(\varepsilon) f(z) d\varepsilon$

Note: this works because of the transformation and the differential area under the integral remains the same ($p_\varphi(z) dz = p(\varepsilon) d\varepsilon$)

$= \nabla_\varphi \int p(\varepsilon) f(g(\varepsilon, \varphi)) d\varepsilon = \nabla_\varphi \mathbb{E}_{p(\varepsilon)}[f(g(\varepsilon, \varphi))] = \mathbb{E}_{p(\varepsilon)}[\nabla_\varphi f(g(\varepsilon, \varphi))]$

We can then estimate this using: $\frac{1}{S} \sum_{s=1}^{S} \mathbb{E}_{p(\varepsilon)}[\nabla_\varphi f(g(\varepsilon^{(s)}, \varphi))]$

Or in practice draw samples from the noise distribution and average the gradients.

D_{KL} Non-Negativity:

$-D_{KL} = -\int_x p(x) \log \frac{p(x)}{q(x)} dx = \int_x p(x) \log \frac{q(x)}{p(x)} dx \leq \log \int_x p(x) \frac{q(x)}{p(x)} dx = \log \int_x q(x) dx = \log(1) = 0$

Trick: $\mathbb{E}(\varphi(x)) \leq \varphi \mathbb{E}(x)$ iff φ is concave (log is)

Reparameterization Trick:

(to enable VAE back prop)

Given Gaussian with μ and σ^2

Assume $\exists$ underlying random variable $\epsilon \sim N(0,1)$

$z = \mu + \sigma\epsilon = f(x, \varepsilon, \theta) \Rightarrow$ can take derivative w.r.t μ, σ (tells us how changes in μ, σ affects output for fixed ε)

Advantage of using N:

when modelling $p_\theta(z)$ and $q_\phi(z|x^{(i)})$ with normal distributions with diagonal covariance matrix, it allows drawing easily from the prior & can calculate D_{KL} analytically

Equation for analytical D_{KL} with N:

For any $p(z) = N(\mu_p, \sigma_p^2 I)$ and $q(z) = N(\mu_q, \sigma_q^2 I)$

$\int p(z) \log(q(z)) dz = -\frac{J}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^{J} \log \sigma_{q,j}^2 - \frac{1}{2}\sum_{j=1}^{J} \frac{\sigma_{p,j}^2 + (\mu_{p,j} - \mu_{q,j})^2}{\sigma_{q,j}^2}$

VAE Backprop:

$\mathbb{E}_z[\log p_\theta(x^{(i)}|z)] - D_{KL}(q_\phi(z|x^{(i)}) \| p_\theta(z))$

$= \mathbb{E}_z[\log p_\theta(x^{(i)}|z)] + \mathbb{E}_z[\log \frac{p_\theta(z)}{q_\phi(z|x^{(i)})}] = \mathbb{E}_z[\log \frac{p_\theta(z) \cdot p_\theta(x^{(i)}|z)}{q_\phi(z|x^{(i)})}] = \mathbb{E}_z[\log \frac{p_\theta(x^{(i)}, z)}{q_\phi(z|x^{(i)})}]$

$\nabla_\phi \mathbb{E}_{\varepsilon \sim q_\phi}\left[\log \frac{p_\theta(x^{(i)}, z)}{q_\phi(z|x^{(i)})}\right] = \nabla_\phi \mathbb{E}_{\varepsilon \sim N(0,I)}\left[\log \frac{p_\theta(x, f(x, \epsilon, \theta))}{q_\phi(f(x, \epsilon, \theta)|x)}\right] = \mathbb{E}_\epsilon\left[\nabla_\phi \log \frac{p_\theta(x, f)}{q_\phi(f|x)}\right]$

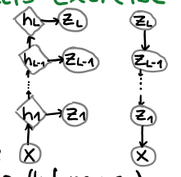
$\approx \frac{1}{k} \sum_{i=1}^{k} \nabla_\phi \log(p_\theta(x, f) / q_\phi(f|x))$

Limitations of VAEs:

Tendency to generate blurry images (believed to be due to injected noise and weak inference models), recently better with inverse autoregressive flow

Hierarchical Latent Variable Models Exercise

To increase expressiveness we make a hierarchy by stacking. Circles are stochastic variables, diamonds are deterministic variables. Equivalent to VAE for $L=1$.



$q_\phi(z_1, \ldots, z_L|x) = q_\phi(z_1|x) \cdot \ldots \cdot q_\phi(z_L|x)$ since the h are all deterministic operations. (Inference)

$p_\theta(z_1, \ldots, z_L) = p_\theta(z_L)\, p(z_{L-1}|z_L) \cdot \ldots \cdot p(z_1|z_2)$ (Generative)

ELBO: $\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x) \| p_\theta(z))$

$= \mathbb{E}_{q_\phi(z_1, \ldots, z_L|x)}[\log p_\theta(x|z_1, \ldots, z_L)] - D_{KL}(q_\phi(z_1, \ldots, z_L|x) \| p_\theta(z_1, \ldots, z_L))$

$= \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z_1, \ldots, z_L|x) \| p_\theta(z_1, \ldots, z_L))$

The last transformation is possible, as x only depends on z_1. We now look at the D_{KL} term:

$= \int_{z_1, \ldots, z_L} q_\phi(z_1, \ldots, z_L|x) \cdot \log\left(\frac{p_\theta(z_1, \ldots, z_L)}{q_\phi(z_1, \ldots, z_L|x)}\right) dz_1, \ldots dz_L$

$= \int_{z_1, \ldots, z_L} \left(\prod_{i=1}^{L} q_\phi(z_i|x)\right) \left(\sum_{i=1}^{L} \log \frac{p_\theta(z_i|z_{i+1})}{q_\phi(z_i|x)}\right) dz_1 \ldots dz_L$ ($p_\theta(z_L|z_{L+1}) = p_\theta(z_L)$)

$= \sum_{i=1}^{L} \int_{z_1, \ldots z_L} \prod_{j=1}^{L} q_\phi(z_j|x) \cdot \log\left(\frac{p_\theta(z_i|z_{i+1})}{q_\phi(z_i|x)}\right) dz_1 \ldots dz_L$

$= \sum_{i=1}^{L} \int_{z_i} \int_{z_{i+1}} \log\left(\frac{p_\theta(z_i|z_{i+1})}{q(z_i|x)}\right) q_\phi(z_i, z_{i+1}|x)\, dz_i\, dz_{i+1}$

$= \int_{z_L} q_\phi(z_L|x) \log\left(\frac{p_\theta(z_L)}{q_\phi(z_L|x)}\right) dz_L + \sum_{i=1}^{L-1} \int_{z_i} \int_{z_{i+1}} q_\phi(z_i, z_{i+1}|x) \log\left(\frac{p_\theta(z_i|z_{i+1})}{q_\phi(z_i|x)}\right) dz_i dz_{i+1}$

$= -D_{KL}(q_\phi(z_L|x) \| p_\theta(z_L)) + \sum_{i=1}^{L-1} \int_{z_{i+1}} q_\phi(z_{i+1}|x) \left[\int_{z_i} q_\phi(z_i|x) \log \frac{p_\theta(z_i|z_{i+1})}{q_\phi(z_i|x)} dz_i\right] dz_{i+1}$

$= -D_{KL}(q_\phi(z_L|x) \| p_\theta(z_L)) - \sum_{i=1}^{L-1} \int_{z_{i+1}} q_\phi(z_{i+1}|x)\, D_{KL}(q_\phi(z_i|x) \| p_\theta(z_i|z_{i+1}))\, dz_{i+1}$

$= -D_{KL}(q_\phi(z_L|x) \| p_\theta(z_L)) - \sum_{i=1}^{L-1} \mathbb{E}_{z_{i+1} \sim q_\phi(z_{i+1}|x)}\left[D_{KL}(q_\phi(z_i|x) \| p_\theta(z_i|z_{i+1}))\right]$

GENERATIVE ADVERSERIAL NETWORKS

GAN Intuition:

Want to sample high dimensional data points from true data distribution (intractable), so draw from simple distribution and use a neural network to transform into output

Use generator $G: \mathbb{R}^D \to X$ and discriminator $D: X \to [0,1]$

GAN Objective:

Consider only the discriminator

$\mathcal{L}(D) = -\frac{1}{N}\left(\sum_{i=1}^{N} (y^i) \cdot \log(D(x^{(i)})) + \sum_{i=1}^{N} (1 - y^i) \log(1 - D(x^i))\right)$

for GANs we know 50% are $y=1$ and 50% are $y=0$.

Our goal is to train D&G jointly:

$= \frac{1}{2}\left(\mathbb{E}_{x \sim p_d}[\log D(x)] + \mathbb{E}_{z \sim p_z}[\log(1 - D(G(z)))]\right)$

$= \frac{1}{2}\left(\mathbb{E}_{x \sim p_d}[\log D(x)] + \mathbb{E}_{x \sim p_g}[\log(1 - D(x))]\right)$

$\Rightarrow$ a good G maximizes loss function of D.

Value function: $V(G,D) = \mathbb{E}_{x \sim p_d}[\log(D(x))] + \mathbb{E}_{x \sim p_g}[\log(1 - D(x))]$

$\Rightarrow G^*$ will fool any D: $G^* = \operatorname{argmin}_G \max_D V(G,D)$

Optimum of $V(G,D)$ iff $p_d = p_g$: Let $D_G^* = \operatorname{argmax}_D V(G,D)$ then the objective becomes $G^* = \operatorname{argmin}_G V(G, D_G^*)$.

If $p_d = p_g$ then $D_G^* = \frac{p_d}{p_d + p_g}$. Proof: $V(G,D) = \int_x p_d(x) \log(D(x)) + \int_z p_z(z) \log(1 - D(G(z)))\, dz = \int_x p_d(x) \log(D(x)) + p_G(x) \log(1 - D(x)) dx$

Optimal D: $\forall a, b \in \{0,0\} \ f(y) = a \cdot \log(y) + b \cdot \log(1-y)$

$f'(x) = 0 \Leftrightarrow \frac{a}{y} - \frac{b}{1-y} = 0 \Rightarrow y = \frac{a}{a+b}$. $f''\left(\frac{a}{a+b}\right) = -\frac{a}{y^2} - \frac{b}{(1-y)^2} < 0 \Rightarrow D_G^* = \frac{p_d}{p_d + p_g}$

Obj. min. JS: $V(D_G^*, G) = \mathbb{E}_{x \sim p_d}[\log \frac{p_d}{p_d + p_g}] + \mathbb{E}_{x \sim p_g}[\log(1 - \frac{p_d}{p_d + p_g})]$

Global min is achieved if $p_d = p_g$ and at $\min V(\cdot,\cdot) = -\log 4$. Proof: $\mathbb{E}_{x \sim p_d}[\log \frac{p_d}{p_d + p_g} \cdot \frac{2}{2}] + \mathbb{E}_{x \sim p_g}[\log \frac{p_g}{p_d + p_g} \cdot \frac{2}{2}]$

$= -\log(2) + \mathbb{E}[\log(\frac{2 \cdot p_d}{p_d + p_g})] - \log(2) + \mathbb{E}[\log(\frac{2 \cdot p_g}{p_d + p_g})]$

$= -\log(4) + D_{KL}(p_d \| \frac{p_d + p_g}{2}) + D_{KL}(p_g \| \frac{p_d + p_g}{2})$ iff $p_d = p_g$ then $= 0$

$= -\log(4) + 2 \cdot D_{JS}(p_d \| p_g)$. Note: last term ≥ 0

Obj:

$\max_G \min_D \mathbb{E}_{x \sim p_{data}}[\log D(x)] + \mathbb{E}_{z \sim p_z}[\log(1 - D(G(z)))]$

GAN Training:

$\min_{\theta_d} \max_{\theta_g} \mathbb{E}_{x \sim p(x)}[\log D_{\theta_d}(x)] + \mathbb{E}_{z \sim p(z)}[\log(1 - D_{\theta_d}(G_{\theta_g}(z)))]$

Alternate: 1) $\max_{\theta_d} \mathbb{E}_{x \sim p(x)}[\log D_{\theta_d}(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D_{\theta_d}(G_{\theta_g}(z)))]$

(both GD) 2) $\min_{\theta_g} \mathbb{E}_{z \sim p_z(z)}[\log(1 - D_{\theta_d}(G_{\theta_g}(z)))]$

GAN Training Problems:

In this problematic region G produces bad samples and gradient is very flat. Can lead to overtraining of the discriminator (G cannot catch up)

Solution: In 2. gradient ascent: $\max_{\theta_g} \mathbb{E}_{z \sim p_z(z)}[\log(D_{\theta_d}(G_{\theta_g}(z)))]$

Generally: need to find Nash-equilibrium in two-player game. This leads to: mode collapse (no sample diversity). happens if there is a saddle in dual energy landscape (the generator learns only variations of a single sample).

Issues with JS divergence: correlates badly with sample quality (do not know when to stop training), D tends to be optimal, JSD saturates $\to$ bad gradients

- solution: use Wasserstein distance

GAN Pseudocode

for #training iterations do
  for k steps do
    sample minibatch of m noise samples $\{z \sim p_g(z)\}$
    sample minibatch of examples $\{x\}$ from data gen. dist.
    update discriminator by ascending on its stochastic grad: $\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} [\log(D(x^{(i)})) + \log(1 - D(G(z)))]$
  sample minibatch of m noise samples $\{z \sim p_g(z)\}$
  update the generator (ascent): $\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log(D(G(z)))$

GAN: Comparison to VAE:

No variational bound is needed, only samples. Generally sharper images. GANs are asymptotically consistent (not proven for VAE)

DC-GAN:

G is an up-sampling network with deconvolutions, D is a CNN. Replace pooling by strided convolutions in D and strided deconvolutions in G. Use batch norm for both. Remove fully connected hidden layers, use ReLU in G except tanh for output, use LeakyReLU in D for all layers. Can do "vector math": smiling ♀ - neutral ♀ + neutral ♂ = smiling ♂

Pix2Pix:

$\mathcal{L}(G,D) = \mathcal{L}_{cGAN}(G,D) + \lambda \mathcal{L}_{L1}(G)$ where

$\mathcal{L}_{cGAN}(G,D) = \mathbb{E}_{x,y}[\log D(x,y)] + \mathbb{E}_{x,z}[1 - \log D(x, G(x,z))]$

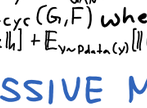
and $\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x,z)\|_1]$

Limitations: requires paired images as training data

Cycle GAN:

$\mathcal{L}(G, F, D_x, D_y) = \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(F, D_X, Y, X) + \lambda \mathcal{L}_{cyc}(G, F)$ where

$\mathcal{L}_{cyc} = \mathbb{E}_{x \sim p_{data}(x)}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p_{data}(y)}[\|G(F(y)) - y\|_1]$



AUTOREGRESSIVE MODELS

Autoregression:

for timeseries, $x_t = b_0 + b_1 x_{t-1} + b_2 x_{t-2}$ taking observations from previous timesteps as input.

i.e. sequence models

Tabular approach: $p(x) = \prod_{i=1}^{n} p(x_i | x_1, \ldots, x_{i-1})$ #params 2^{n-1}

can also specify $p(x_i | x_1, \ldots, x_{i-1}) = \text{Ber}(f(x_1, \ldots, x_{i-1}))$

Fully visible belief networks: $\hat{x}_i = p(x_i = 1 \mid x_1, \ldots, x_{i-1})$ modeled via logistic regression: $f(x_1, \ldots, x_{i-1}) = \sigma(\alpha_0^i + \ldots + \alpha_{i-1}^i x_{i-1})$, has $O(n^2)$ parameters

Neural Autoregressive Density Estimator (NADE)

autoencoder-like network to learn $p(x_i = 1 | x_1, \ldots, x_{i-1})$

$h^{(i)} = \sigma(b + W_{:,<i} x_{<i})$ (first i cols), $\hat{x}_i = \sigma(c_i + V_{i,:} h^{(i)})$ (i-th row) for binary data

(we can use that $(b + W_{:,<i+1} x_{<i+1}) - (b + W_{:,<i} x_{<i}) = W_{:,i} x_i$)

Train by max. $\frac{1}{T}\sum_{t=1}^{T} \log(p(x^{(t)})) = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n} \log p(x_i^{(t)} | x_{<i}^{(t)})$

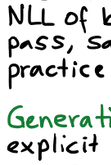
Advantages: efficient (computations in $O(T)$),

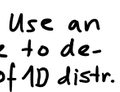
could make use of second order optimizers, easily extendable to other types of observations (reals, multinomials). Random input order is fine. Training: take data from training, Inference: sampling
Extensions: RNADE: conditionals modelled by mixture of Ganssians, DeepNADE: a DNN trained to assign a cond. distr. to any given var. given any subset of others (order-less), ConvNADE

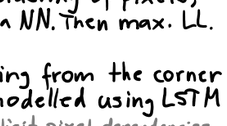
Masked Autoencoder Distribution Estimator (MADE)
Constrain autoencoder s.t. output can be used as conditional $p(x_i|x_{<i})$. To fulfill autoregressive property: no computational path output unit x_d and any of the input units $x_d, \dots, x_D$ must exist. Can also be seen as masks $m_i^{(l)}$: 1 if unit i of layer l is connected to unit j of layer l-1. Train with NLL of binary x. Computing p(x) is just forward pass, sampling requires D forward passes. In practice very large hidden layers necessary.

Generative Models for Natural Images: Use an explicit density model & use chain rule to decompose likelihood of image x into prod. of 1D distr. $p(x) = \prod_{i=1}^{n} p(x_i|x_{<i})$ → need to define ordering of pixels, complex distributions → paramet. via NN. Then max. LL.

PixelRNN: Generate pixel starting from the corner and dependency to prev. pixels modelled using LSTM
Issue: sequential gen s owl due to explicit pixel dependencies

PixelCNN: Still generate starting from corner, but model the dependency as CNN:  (masked convolutions)
$p(x_i|x_{<i}) = p(x_{i,R}|x_{<i}) \cdot p(x_{i,G}|x_{i,R}, x_{<i}) \cdot p(x_{i,B}|x_{i,G}, x_{i,R}, x_{<i})$

stacks can be used to speed it up. Add information between vertical and horizontal stack (preserve pixel dependencies)
$h_{k+1} = \tanh(W_{k,f} * h_k) \odot \sigma(W_{k,g} * h_k)$. Training is faster than PixelRNN (can parallelize convolutions) but still slow (sequential generation).
Advantages: explicit likelihood, likelihood of training data is good eval. metric, good samples.

Dilated convolution: Receptive field size can increase exponentially with dilated layers. 

WaveNet: Temporal convolutional networks uses dilated convolution. 
↳cannot use strided conv, due to need to preserve resolution

Variational RNN (VRNN): RNNs are autoregressive models Increase expressive power of RNNs by incorporating stochastic latent variables into hidden state of RNN Include one VAE per timestep. Including a dynamic prior explicitly models temporal dependencies between timesteps. Can include conditional (e.g. style and content of handwriting.

Stochastic Temporal Convolutional Networks (STCN)
Combines computational advantages of TCNs with expressiveness of stochastic RNNs: integrate a hierarchical VAE with a TCN.

EYE GAZE ESTIMATION

Eye Gaze Estimation: Challenges: head poses, gaze directions, illumination conditions, personal appearances. Applications: explicit eye input (eye typing, gaming), user modelling (What are they doing/focusing on), attentive user interface (autopause), passive eye monitoring (ads).

Typical Gaze Estimation Pipeline: Grab frame, segment face, detect landmarks, segment eye, 3D gaze direction or 2D point of regard/gaze algorithm

Geometry-based Methods: hand-picked features (e.g. iris center-eye-vector) often with 3D head pose feature, not robust to head movements, eyeball model fitting calibration from few samples, robust to head movement.

LeNet/GazeNet: uses image, DNN. Can also include both eyes, face & face grid.

Synthetic data: need domain adaption (network that takes unlabeled real images & refines them).

CLASSICAL REINFORCEMENT LEARNING

Reinforcement Learning: Uses: games, robot control, attention based models, logistics & operations
Goal: achieve high cumulative reward over time.
Value function of policy π, V_π, expresses expected cumulative reward for each state when following π.
$V_\pi(s) = \mathbb{E}_\pi[G_t|S_t=s] = \mathbb{E}_\pi[R_{t+1} + \gamma G_t | S_t = s] = \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a)[r + \gamma \mathbb{E}_\pi[G_{t+1}|S_{t+1}=s']]$
$= \sum_a \pi(a,s) \sum_{s'} \sum_r p(s', r|s, a)[r + \gamma V_\pi(s')]$ where $G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$

MDP: states S, actions A, reward function r, transition function p, initial state s_0, discount factor γ

Value Iteration: Given a value function V_π, we can derive a greedy policy π':
$\pi'(s) = \operatorname{argmax}_{a \in A}(r(s,a) + \gamma V_\pi(p(s,a)))$. One can show $V_{\pi'}(s) \geq V_\pi(s)$. The optimal policy π^* is greedy w.r.t. V_{π^*}. V_{π^*} fulfills the Bellman Optimality Equation: $V^*(s) = \max_{a \in A}\{r(s,a) + \gamma V^*(p(s,a))\}$. Can use step cost instead of γ. Update backwards each time doing the action that gives the highest reward until nothing changes anymore.

```
i ← 0, V_0 ← 0
while Δ > θ
    Δ ← 0
    for all s ∈ S
        v ← V(s)
        V(s) ← max_{a∈A} Σ_{s',r} p(s', r|s, a)[r + γV(s')]
        Δ ← max(Δ, |v − V(s)|)
π_i ← greedy policy w.r.t. V_i
```

Theorem: value iteration converges. At convergence found the optimal V^*. Need to find the optimal (stationary & greedy) policy.

Advantages of DP: Exact methods, policy/value iteration guaranteed to converge in finite number of iterations, value iteration typically more efficient easy implementation

Disadvantages of DP: need to know transition probability matrix, need to iterate over the whole state space (very expensive), requires memory proportional to size of state space, generally computationally expensive

Problem with large state space: we will never visit all states, but only a small subset. We cannot estimate the values of these states.

Monte Carlo Methods: value function is $\mathbb{E}[\cdot]$: $V_\pi(s) = \mathbb{E}_{a_t \sim \pi(s_t), s_{t+1} \sim p(s_t, a_t)}[\sum_{i=0} \gamma^i r(s_{t+1}, a_{t+1})]$, so we can just run policies with the given policy and compute samples of: $\sum_{i=0} \gamma^i r(s_{t+1}, a_{t+1})$
(follow policy, update values, update policy, repeat)
Advantages: Unbiased estimate, do not need to know system dynamics. Disadvantages: High variance, exploration/exploitation dilemma, need termination state (slow for long episodes)

Temporal Difference Learning: For each step with action a from s to s' that we take with our policy, we compute the difference to our current estimate and update our value function:
$\Delta V(s) = r(s,a) + \gamma V(s') - V(s)$, $V(s) \leftarrow V(s) + \alpha \Delta V(s)$, $\alpha > 0$
SARSA: On policy: Computes the Q-Value according to a policy and then the agent follows that policy.
Q-Learning: Off policy: Computes the Q-Value according to a greedy policy, but the agent follows a different exploration policy:
$\Delta Q(S,A) = R_{t+1} + \gamma \max_a\{Q(S',a)\} - Q(S,A)$, $Q(S,A) \leftarrow Q(S,A) + \alpha \Delta Q(S,A)$
off policy since π for updating Q different from π' that we use for data collection.
Advantages: Less variance than Monte Carlo sampling due to bootstrapping. More sample efficient, do not need to know the transition probability matrix. Disadvantages: biased due to bootstrapping, exploration/exploitation dilemma can behave poorly in stochastic environments. requires exploring full state-action space (does not work for continuous action space)

Taxonomy of RL methods
Policy Optimization → Grad-free methods -Param. Perturbation -Evolutionary Strat.
Policy Optimization → Policy grad.
Methods based on TD learning → Policy iter. → Actor-critic methods
Methods based on TD learning → Value iter. → Q-learning

DEEP REINFORCEMENT LEARNING

Policy gradients: $\pi(a_t|s_t) = \mathcal{N}(\mu_t, \sigma_t^2|s_t)$. We can collect rollout trajectories: $p(\tau) = p(s_1, a_1, \dots, s_T, a_T) = p(s_1) \prod_{t=1}^{T} \pi(a_t|s_t) p(s_{t+1}|a_t, s_t)$ then we update the policy: good traj. more & bad traj. less likely.

Difference Policy optimization vs DP: directly optimize desired quantity (not indirect & exploit problem structure and self-consistency), more compatible with modern ML machinery, more versatile & flexible, more compatible with auxiliary objectives, less compatible with off-policy and exploration, less sample efficient.

Policy gradient: goal: $\Theta^* = \operatorname{argmax} J(\theta)$ by updating $\Theta \leftarrow \Theta + \nabla_\theta J(\theta)$ with gradient ascent.
$J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)}[\sum_t \gamma^t r(s_t, a_t)] = \mathbb{E}_{\tau \sim p_\theta(\tau)}[r(\tau)] = \int p(\tau) r(\tau) d\tau$
$\nabla_\theta J(\theta) = \int \nabla_\theta p(\tau) r(\tau) d\tau = \int \frac{p(\tau)}{p(\tau)} \nabla_\theta p(\tau) r(\tau) d\tau = \int p(\tau) \frac{\nabla p(\tau)}{p(\tau)} r(\tau) d\tau$
$= \int p(\tau) \nabla_\theta \log p(\tau) r(\tau) d\tau$ (since $\frac{d \log u(x)}{dx} = \frac{u'(x)}{u}$) $= \mathbb{E}_{\tau \sim p(\tau)}[\nabla_\theta \log p(\tau) r(\tau)]$
$\log p(\tau) = \log p(s_1, a_1, \dots, s_T, a_T) = \log[p(s_1) \prod_{t=0}^{T} \pi_\theta(a_t|s_t) p(s_{t+1}|a_t, s_t)]$
$\Rightarrow \nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p(\tau)}[(\sum_{t=0}^{T} \log \pi_\theta(a_t|s_t))(\sum_{t=0}^{T} \gamma^t r(s_t, a_t))]$
Valid even if reward is discontinuous or unknown sample space of path is a discrete set.
Gradient tries to increase probability with pos. r and decrease probability with small/neg. r.

Reinforce Trick
$\nabla_\theta p_\theta(x) = p_\theta(x) \nabla_\theta \log p_\theta(x)$ which allows us:
$\nabla_\theta \mathbb{E}_{x \sim p_\theta(x)}[f(x)] = \nabla_\theta \int f(x) p_\theta(x) dx = \int f(x) \nabla_\theta p_\theta(x) dx$
$= \int f(x) p_\theta(x) \nabla_\theta \log p_\theta(x) dx = \mathbb{E}_{x \sim p_\theta(x)}[f(x) \nabla_\theta \log p_\theta(x)]$

Reinforce: sample trajectories τ_i by rolling out the policy, i.e. sampling a random action from the current policy until the end of the episode. Monte Carlo type of RL method that evaluates full trajectories (episodes to update policy
$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_i (\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t^i|s_t^i))(\sum_{t=0}^{T} \gamma^t r(s_t^i, a_t^i))$, $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$
(unbiased estimate) Variance problem: this produces noisy gradients: If all cumulative rewards are positive, we make all trajectories more likely. If all traj. rewards are 0, we might decrease probability of good trajectories.
Baseline: To reduce variance, we can introduce a baseline $b(s_t^i)$ (produces okay trajectories, actually any function that does not depend on a_t, e.g. average reward or estimate of state-value function)
$r(\tau) = \sum_{t=0}^{T} \gamma^t r(s_t^i, a_t^i) - b(s_t^i)$. Still unbiased!

Actor-critic: Use bootstrapping for reward estimation:
$\nabla_\theta J(\theta) = \frac{1}{N} \sum_i \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t^i|s_t^i)(r(s_t^i, a_t^i) + \gamma V_\phi(s_{t+1}^i) - V_\phi(s_t^i))$
Both π and V are represented by neural nets.

TRPO: tries to solve stepsize problem (slow learning vs. disastrous effects) by formulating it as a constraint optimization problem. (second order optimization, computationally heavy).

PPO: Relax hard TRPO constraint to penalize with tuning parameter (allows some bad decisions but generally within trust region). Even better: clipped objective. Simpler, computationally easier than TRPO.

DDPG: When selecting action we add noise to it. Also uses experience replay & target networks. (off-policy)

Issues with DRL: Exploration vs Exploitation (get stuck in local minima), Reward hacking (just get power-ups), Sample efficiency (need excessive amounts of data, infeasible in real-world scenarios), domain knowledge (small tricks lead to different results e.g. comparison of different implementations)