This document is a summary of the *Machine Perception* course at ETH Zürich. This summary was created during the spring semester of 2024. Due to updates to the syllabus content, some material may no longer be relevant for future versions of the course. I do not guarantee correctness or completeness, nor is this document endorsed by the lecturers. The order of the chapters is not necessarily the order in which they were presented in the course. For the full LaTeX source code, visit `github.com/Jovvik/eth-cheatsheets`.
All figures are created by the author, and, assuming the rules have not been changed, are not allowed to be used as a part of a cheat sheet during the exam.

1 CNN

T is linear if $T(\alpha\mathbf{u}+\beta\mathbf{v}) = \alpha T(\mathbf{u}) + \beta T(\mathbf{v})$, invariant to f if $T(f(\mathbf{u})) = T(\mathbf{u})$, equivariant to f if $T(f(\mathbf{u})) = f(T(\mathbf{u}))$. Any linear shift-equivariant T can be written as a convolution. Convolution: $I'(i,j) = \sum_{m=-k}^{k}\sum_{n=-k}^{k} K(-m,-n)I(m+i,n+j)$. Correlation: $I'(i,j) = \sum_{m=-k}^{k}\sum_{n=-k}^{k} K(m,n)I(m+i,n+j)$. Conv. forward: $z^{(l)} = w^{(l)} * z^{(l-1)} + b^{(l)} = \sum_m \sum_n w_{m,n}^{(l)} z_{i-m,j-n}^{(l-1)} + b^{(l)}$. Backward inputs: $\delta^{(l-1)} = \frac{\partial C}{\partial z_{i,j}^{(l-1)}} = \delta^{(l)} * \mathrm{ROT}_{180}(w^{(l)})$, backward kernel: $\frac{\partial C}{\partial w_{m,n}^{(l)}} = \delta^{(l)} * \mathrm{ROT}_{180}(z^{(l-1)})$. Width or height after conv or pool: (in+2·pad−dil·(kern−1)−1)/stride+1, rounded down. Channels = number of kernels.

1D conv as matmul: $\begin{bmatrix} k_1 & 0 & \dots & 0 \\ k_2 & k_1 & & \vdots \\ k_3 & k_2 & k_1 & 0 \\ 0 & k_3 & k_2 & 0 \\ \vdots & \vdots & \ddots & \vdots \end{bmatrix}$

Backprop example (rotate K):

$$\underset{\mathbf{X}}{\begin{bmatrix} 0&0&1&1&0\\0&1&0&1&1\\1&1&1&1&0\\1&0&1&1&0\\1&1&0&1&1 \end{bmatrix}} \rightarrow \underset{\mathbf{K}}{\begin{bmatrix} 1&0&1\\0&1&0\\0&1&0 \end{bmatrix}} \rightarrow \underset{\mathbf{Y}=\mathbf{X}*\mathbf{K}}{\begin{bmatrix} 3&3&3\\4&2&3\\2&3&3 \end{bmatrix}} \rightarrow$$

$$\rightarrow \underset{\mathbf{Y}'=\mathrm{Pool}(\mathbf{Y})}{[4]} \;\Big|\; \underset{\partial E/\partial \mathbf{Y}'}{[1]} \rightarrow \underset{\partial E/\partial \mathbf{Y}}{\begin{bmatrix} 0&0&0\\1&0&0\\0&0&0 \end{bmatrix}} \rightarrow$$

$$\rightarrow \underset{\partial E/\partial \mathbf{K}}{\begin{bmatrix} 1&0&1\\1&1&1\\0&1&0 \end{bmatrix}} \rightarrow \underset{\partial E/\partial \mathbf{V}}{\begin{bmatrix} 0&0&0&0&0\\0&1&0&0&0\\0&1&0&0&0\\1&0&1&0&0\\0&0&0&0&0 \end{bmatrix}}$$

Max-pooling: $z^{(l)} = \max z_i^{(l-1)}$. $i^* := \arg\max_i z_i^{(l-1)}$, $\frac{\partial z^{(l)}}{\partial z_i^{(l-1)}} = [i = i^*]$, $\delta^{(l-1)} = \delta_{i^*}^{(l)}$.

Unpooling: nearest-neighbor (duplicate), bed of nails (only top left, rest 0), max-unpooling (remember where max came from when pooling). Learnable upsampling: transposed conv, output is copies of filter weighted by input, summed on overlaps.

Sizes input: (C_{in}, H_{in}, W_{in}), output: $(C_{out}, H_{out}, W_{out})$. Kernel: $k_1 \times k_2$, padding: $p_1 \times p_2$, stride: $s_1 \times s_2$, dilation: $d_1 \times d_2$. $H_{out} = (H_{in} + 2p_1 - d_1(k_1-1) - 1)s_1^{-1} + 1$, $W_{out} = (W_{in} + 2p_2 - d_2(k_2-1) - 1)s_2^{-1} + 1$.

2 RNN

Vanilla RNN: $\hat{y}_t = \mathbf{W}_{hy}\mathbf{h}_t$, $\mathbf{h}_t = f(\mathbf{h}_{t-1}, \mathbf{x}_t, \mathbf{W})$, usually $\mathbf{h}_t = \tanh(\mathbf{W}_{hh}\mathbf{h}_{t-1} + \mathbf{W}_{xh}\mathbf{x}_t)$.

BPTT: $\frac{\partial L}{\partial \mathbf{W}} = \sum_t \frac{\partial L_t}{\partial \mathbf{W}}$, treat unrolled model as multi-layer. $\frac{\partial L_t}{\partial W}$ has a term of $\frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_k} = \prod_{i=k+1}^{t} \frac{\partial \mathbf{h}_i}{\partial \mathbf{h}_{i-1}} = \prod_{i=k+1}^{t} \mathbf{W}_{hh}^{\top} \operatorname{diag} f'(\mathbf{h}_{i-1})$.

Exploding/vanishing gradients: $\mathbf{h}_t = \mathbf{W}^t \mathbf{h}_1$. If $\mathbf{W}$ is diagonaliz., $\mathbf{W} = \mathbf{Q} \operatorname{diag} \boldsymbol{\lambda} \mathbf{Q}^{\top} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{\top}, \mathbf{Q}\mathbf{Q}^{\top} = \mathbf{I} \Rightarrow \mathbf{h}_t = (\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{\top})^t \mathbf{h}_1 = (\mathbf{Q}(\operatorname{diag}\boldsymbol{\lambda})^t\mathbf{Q}^{\top})\mathbf{h}_1 \Rightarrow \mathbf{h}_t$ becomes the dominant eigenvector of $\mathbf{W}$. $\frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_k}$ has this issue. Long-term contributions vanish, too sensitive to recent distrations. **Truncated BPTT**: take the sum only over the last κ steps. **Gradient clipping** $\frac{\text{threshold}}{\|\nabla\|}\nabla$ fights exploding gradients.

2.1 LSTM

We want constant error flow, not multiplied by W^t.

- Input gate: which values to write,
- forget gate: which values to reset,
- output gate: which values to read,
- gate: candidate values to write to state.

$$\begin{pmatrix} \mathbf{i} \\ \mathbf{f} \\ \mathbf{o} \\ \mathbf{g} \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} W \begin{pmatrix} \mathbf{x}_t \\ \mathbf{h}_{t-1} \end{pmatrix}$$

$$\mathbf{c}_t = \mathbf{f} \odot \mathbf{c}_{t-1} + \mathbf{i} \odot \mathbf{g}$$

$$\mathbf{h}_t = \mathbf{o} \odot \tanh(\mathbf{c}_t)$$

3 Generative modelling

Learn $p_{\text{model}} \approx p_{\text{data}}$, sample from p_{model}.

- Explicit density:
  - Approximate:
    - Variational: VAE, Diffusion
    - Markov Chain: Boltzmann machine
  - Tractable:
    - Autoregressive: FVSBN/NADE/MADE, Pixel(C/R)NN, WaveNet/TCN, Autor. Transf.,
    - Normalizing Flows
- Implicit density:
  - Direct: Generative Adversarial Networks
  - MC: Generative Stochastic Networks

Autoencoder: $X \to Z \to X$, $g \circ f \approx$ id, f and g are NNs. Optimal linear autoencoder is PCA. Undercomplete: $|Z| < |X|$, else overcomplete. Overcomp. is for denoising, inpainting. Latent space should be continuious and interpolable. Autoencoder spaces are neither, so they are only good for reconstruction.

4 Variational AutoEncoder (VAE)

Sample z from prior $p_\theta(z)$, to decode use conditional $p_\theta(x \mid z)$ defined by a NN. $D_{\mathrm{KL}}(P\|Q) := \int_x p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x$: KL divergence, measure similarity of prob. distr. $D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$, $D_{\mathrm{KL}}(P\|Q) \geq 0$

z can also be categorical. Likelihood $p_\theta(x) = \int_z p_\theta(x \mid z) p_\theta(z) \mathrm{d}z$ is hard to maximize, let encoder NN be $q_\phi(z \mid x)$, $\log p_\theta(x^i) = \mathbb{E}_z\left[\log p_\theta(x^i \mid z)\right] - D_{\mathrm{KL}}(q_\phi(z \mid x^i) \| p_\theta(z)) + D_{\mathrm{KL}}(q_\phi(z \mid x^i) \| p_\theta(z \mid x^i))$. Red is intractable, use ≥ 0 to ignore it; Orange is reconstruction loss, clusters similar samples; Purple makes posterior close to prior, adds cont. and interp. Orange − Purple is **ELBO**, maximize it.

$$x \xrightarrow{\text{enc}} \mu_{z|x}, \Sigma_{z|x} \xrightarrow{\text{sample}} z \xrightarrow{\text{dec}} \mu_{x|z}, \Sigma_{x|z} \xrightarrow{\text{sample}} \hat{x}$$

Backprop through sample by reparametr.: $z = \mu + \sigma\epsilon$. For inference, use μ directly.

Disentanglement: features should correspond to distinct factors of variation. Can be done with semi-supervised learning by making z conditionally independent of given features y.

4.1 β-VAE

$\max_{\theta,\phi} \mathbb{E}_x\left[\mathbb{E}_{z \sim q_\phi} \log p_\theta(x \mid z)\right]$ to disentangle s.t. $D_{\mathrm{KL}}(q_\phi(z \mid x) \| p_\theta(z)) < \delta$, with KKT: max Orange − βPurple.

5 Autoregressive generative models

Autoregression: use data from the same input variable at previous time steps

Discriminative: $P(Y \mid X)$, generative: $P(X, Y)$, maybe with Y missing. Sequence models are generative: from $x_i \dots x_{i+k}$ predict x_{i+k+1}.

Tabular approach: $p(\mathbf{x}) = \prod_i p(x_i \mid \mathbf{x}_{<i})$, needs 2^{i-1} params. Independence assumption is too strong. Let $p_{\theta_i}(x_i \mid \mathbf{x}_{<i}) = \mathrm{Bern}(f_i(\mathbf{x}_{<i}))$, where f_i is a NN. **Fully Visible Sigmoid Belief Networks**: $f_i = \sigma(\alpha_0^{(i)} + \boldsymbol{\alpha}^{(i)}\mathbf{x}_{<i}^{\top})$, complexity n^2, but model is linear.

Neural Autoregressive Density Estimator: add hidden layer. $\mathbf{h}_i = \sigma(\mathbf{b} + \mathbf{W}_{\cdot,<i}\mathbf{x}_{<i})$, $\hat{x}_i = \sigma(c_i + \mathbf{V}_{i,\cdot}\mathbf{h}_i)$. Order of $\mathbf{x}$ can be arbitrary but fixed. Train by max log-likelihood in $\mathcal{O}(TD)$, can use 2nd order optimizers, can use **teacher forcing**: feed GT as previous output.

Extensions: Convolutional; Real-valued: conditionals by mixture of gaussians; Order-less and deep: one DNN predicts $p(x_k \mid x_{i_1} \dots x_{i_j})$.

Masked Autoencoder Distribution Estimator: mask out weights s.t. no information flows from $x_d \dots$ to $\hat{x}_d$. Large hidden layers needed. Trains as fast as autoencoders, but sampling needs D forward passes.

PixelRNN: generate pixels from corner, dependency on previous pixels is by RNN (LSTM). **PixelCNN**: also from corner, but condition by CNN over context region (perceptive field) $\Rightarrow$ parallelize. For conditionals use masked convolutions. Channels: model R from context, G from R + cont., B from G + R + cont. Training is parallel, but inference is sequential $\Rightarrow$ slow. Use conv. stacks to mask correctly.

NLL is a natural metric for autoreg. models, hard to evaluate others.

WaveNet: audio is high-dimensional. Use dilated convolutions to increase perceptive field with multiple layers.

AR does not work for high res images/video, convert the images into a series of tokens with an AE: Vector-quantized VAE. The codebook is a set of vectors. $x \xrightarrow{\text{enc}} z \xrightarrow{\text{codebook}} z_q \xrightarrow{\text{dec}} \hat{x}$. We can run an AR model in the latent space.

5.1 Attention

$\mathbf{x}_t$ is a convex combination of the past steps, with access to all past steps. For $X \in \mathbb{R}^{T \times D}$: $K = XW_K, V = XW_V, Q = XW_Q$. Check pairwise similarity between query and keys via dot product: let attention weights be $\boldsymbol{\alpha} = \mathrm{Softmax}(QK^{\top}/\sqrt{D})$, $\boldsymbol{\alpha} \in \mathbb{R}^{1 \times T}$. Adding mask M to avoid looking into the future:

$$X = \mathrm{Softmax}\left(\frac{(XW_Q)(XW_K)^{\top}}{\sqrt{D}} + M\right)(XW_V)$$

Multi-head attn. splits W into h heads, then concatenates them. Positional encoding injects information about the position of the token. Attn. is $\mathcal{O}(T^2D)$.

6 Normalizing Flows

VAs dont have a tractable likelihood, AR models have no latent space. Want both. Change of variable for $x = f(z)$: $p_x(x) = p_z(f^{-1}(x))\left|\det \frac{\partial f^{-1}(x)}{\partial x}\right| = p_z(f^{-1}(x))\left|\det \frac{\partial f(z)}{\partial z}\right|^{-1}$. Map $Z \to X$ with a deterministic invertible f_θ. This can be a NN, but computing the determinant is $O(n^3)$. If the Jacobian is triangular, the determinant is $O(n)$. To do this, add a coupling layer: $\begin{pmatrix} y^A \\ y^B \end{pmatrix} = \begin{pmatrix} h(x^A, \beta(x^B)) \\ x^B \end{pmatrix}$, where β is any model, and h is elementwise.

$$\begin{pmatrix} x^A \\ x^B \end{pmatrix} = \begin{pmatrix} h^{-1}(y^A, \beta(y^B)) \\ y^B \end{pmatrix}, J = \begin{pmatrix} h' & h'\beta' \\ 0 & 1 \end{pmatrix}$$

Stack these for expressivity, $f = f_k \circ \dots f_1$. $p_x(x) = p_z(f^{-1}(x)) \prod_k \left|\det \frac{\partial f_k^{-1}(x)}{\partial x}\right|$.
Sample $z \sim p_z$ and get $x = f(z)$.
- Squeeze: reshape, increase chan.
- ActNorm: batchnorm with init. s.t. output $\sim \mathcal{N}(0, \mathbf{I})$ for first minibatch. $\mathbf{y}_{i,j} = \mathbf{s} \odot \mathbf{x}_{i,j} + \mathbf{b}$, $\mathbf{x}_{i,j} = (\mathbf{y}_{i,j} - \mathbf{b})/\mathbf{s}$, $\log\det = H \cdot W \cdot \sum_i \log|\mathbf{s}_i|$: linear.
- 1×1 conv: permutation along channel dim. Init $\mathbf{W}$ as rand. ortogonal $\in \mathbb{R}^{C \times C}$ with $\det \mathbf{W} = 1$. $\log\det = H \cdot W \cdot \log|\det \mathbf{W}|$: $O(C^3)$. Faster: $\mathbf{W} := \mathbf{PL}(\mathbf{U} + \mathrm{diag}(s))$, where $\mathbf{P}$ is a random fixed permut. matrix, $\mathbf{L}$ is lower triang. with 1s on diag., $\mathbf{U}$ is upper triang. with 0s on diag., $\mathbf{s}$ is a vector. Then $\log\det = \sum_i \log|\mathbf{s}_i|$: $O(C)$ Conditional coupling: add parameter $\mathbf{w}$ to β.

SRFlow: use flows to generate many high-res images from a low-res one. Adds affine injector between conv. and coupling layers. $\mathbf{h}^{n+1} = \exp(\beta^n_{\theta,s}(\mathbf{u})) \cdot \mathbf{h}^n + \beta_{\theta,b}(\mathbf{u})$, $\mathbf{h}^n = \exp(-\beta^n_{\theta,s}(\mathbf{u})) \cdot (\mathbf{h}^{n+1} - \beta^n_{\theta,b}(\mathbf{u}))$, $\log\det = \sum_{i,j,k} \beta^n_{\theta,s}(\mathbf{u}_{i,j,k})$.
StyleFlow: Take StyleGAN and replace the network $\mathbf{z} \to \mathbf{w}$ (aux. latent space) with a normalizing flow conditioned on attributes.
C-Flow: condition on other normalizing flows: multimodal flows. Encode original image $\mathbf{x}^1_B$: $\mathbf{z}^1_B = f^{-1}_\phi(\mathbf{x}^1_B \mid \mathbf{x}^1_A)$; encode extra info (image, segm. map, etc.) $\mathbf{x}^2_A$: $\mathbf{z}^2_A = g^{-1}_\theta(\mathbf{x}^2_A)$; generate new image $\mathbf{x}^2_B$: $\mathbf{x}^2_B = f_\phi(\mathbf{z}^1_B \mid \mathbf{z}^2_A)$.
Flows are expensive for training and low res. The latent distr. of a flow needn't be $\mathcal{N}$.

7 Generative Adversarial Networks (GANs)

Log-likelihood is not a good metric. We can have high likelihood with poor quality by mixing in noise and not losing much likelihood; or low likelihood with good quality by remembering input data and having sharp peaks there.
Generator $G : \mathbb{R}^Q \to \mathbb{R}^D$ maps noise z to data, **discriminator** $D : \mathbb{R}^D \to [0,1]$ tries to decide if data is real or fake, receiving both gen. outputs and training data. Train D for k steps for each step of G.
Training GANs is a min-max process, which are hard to optimize. $V(G, D) = \mathbb{E}_{\mathbf{x} \sim p_{\mathrm{d}}} \log(D(\mathbf{x})) + \mathbb{E}_{\hat{\mathbf{x}} \sim p_{\mathrm{m}}} \log(1 - D(\hat{\mathbf{x}}))$
For G the opt. $D^* = p_{\mathrm{d}}(\mathbf{x})/(p_{\mathrm{d}}(\mathbf{x}) + p_{\mathrm{m}}(\mathbf{x}))$.
Jensen-Shannon divergence (symmetric): $D_{\mathrm{JS}}(p\|q) = \frac{1}{2} D_{\mathrm{KL}}(p\|\frac{p+q}{2}) + \frac{1}{2} D_{\mathrm{KL}}(p\|\frac{p+q}{2})$.
Global minimum of $D_{\mathrm{JS}}(p_{\mathrm{d}}\|p_{\mathrm{m}})$ is the glob. min. of $V(G,D)$, $V(G, D^*) = -\log(4)$ and at optimum of $V(D^*, G)$ we have $p_d = p_m$.
If G and D have enough capacity, at each update step D reaches D^* and p_{m} improves $V(p_{\mathrm{m}}, D^*) \propto \sup_D \int_{\mathbf{x}} p_{\mathrm{m}}(\mathbf{x}) \log(-D(\mathbf{x})) \mathrm{d}\mathbf{x}$, then $p_{\mathrm{m}} \to p_{\mathrm{d}}$ by convexity of $V(p_{\mathrm{m}}, D^*)$ wrt. p_{m}. These assumptions are too strong.
If D is too strong, G has near zero gradients and doesn't learn ($\log'(1 - D(G(z))) \approx 0$). Use gradient ascent on $\log(D(G(z)))$ instead.
Mode collapse: G only produces one sample or one class of samples. Solution: **unrolling** — use k previous D for each G update.
GANs are hard to compare, as likelihood is intractable. FID is a metric that calculates the distance between feature vectors calculated for real and generated images.
Adding gradient penalty for D stabilizes training. **Profressive Growing** GAN: generate low-res image, then high-res during training.
AdaIN: Similar to attention. $c' = \gamma(s) \odot \frac{c - \mu(c)}{\sigma(c)} + \beta(s)$, where c is the content, s is the style. **StyleGAN**: learn intermediate latent space $\mathcal{W}$ with FCs, batchnorm with scale and mean from $\mathcal{W}$, add noise at each layer using AdaIn.
GAN **inversion**: find z s.t. $G(z) \approx x \Rightarrow$ manipulate images in latent space, inpainting. If G predicts image and segmentation mask, we can use inversion to predict mask for any image, even outside the training distribution.

7.1 3D GANs

3D GAN: voxels instead of pixels. PlatonicGAN: 2D input, 3D output differentiably rendered back to 2D for D.
HoloGAN: 3D GAN + 2D superresolution GAN
GRAF: radiance fields more effic. than voxels
GIRAFFE: GRAF + 2D conv. upscale
EG3D: use 3 2D images from StyleGAN for features, project each 3D point to tri-planes.

7.2 Image Translation

E.g. sketch $X \to$ image Y. Pix2Pix: $G : X \to Y$, $D : X, Y \to [0,1]$. GAN loss $+L_1$ loss between sketch and image. Needs pairs for training.
CycleGAN: unpaired. Two GANs $F : X \to Y$, $G : Y \to X$, cycle–consistency loss $F \circ G \approx \mathrm{id}$; $G \circ F \approx \mathrm{id}$ plus GAN losses for F and G.
BicycleGAN: add noise input.
Vid2vid: video translation.

8 Diffusion models

High quality generations, better diversity, more stable/scalable.
Diffusion (forward) step q: adds noise to $\mathbf{x}_t$ (not learned). Denoising (reverse) step p_θ: removes noise from $\mathbf{x}_t$ (learned).
$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\sqrt{1-\beta}\mathbf{x}_{t-1}, \beta_t \mathbf{I})$
$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mu_\theta(\mathbf{x}_t, t), \sigma_t^2 \mathbf{I})$
β_t is the variance schedule (monotone ↑). Let $\alpha_t := 1 - \beta_t$, $\bar{\alpha}_t := \prod \alpha_i$, then $q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \Rightarrow \mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$.
Denoising is not tractable naively: $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) q(\mathbf{x}_{t-1})/q(\mathbf{x}_t)$, $q(\mathbf{x}_t) = \int q(\mathbf{x}_t \mid \mathbf{x}_0) q(\mathbf{x}_0) \mathrm{d}\mathbf{x}_0$.
Conditioning on $\mathbf{x}_0$ we get a Gaussian. Learn model $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) \approx q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)$ by predicting the mean.
$\log p(\mathbf{x}_0) \geq \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \log\left(\frac{p(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right) = \mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)} \log p_\theta(\mathbf{x}_0|\mathbf{x}_1) - D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0)\|p(\mathbf{x}_T)) - \sum_{t=2}^T \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)} D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)\|p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))$, where orange and purple are the same as in VAEs, and blue are the extra loss functions. In a sense VAEs are 1-step diffusion models.
t-th denoising is just $\arg\min_\theta \frac{1}{2\sigma_q^2(t)} \|\mu_\theta - \mu_q\|_2^2$, so we want $\mu_\theta(\mathbf{x}_t, t) \approx \mu_q(\mathbf{x}_t, \mathbf{x}_0)$. $\mu_q(\mathbf{x}_t, \mathbf{x}_0)$ can be written as $\frac{1}{\sqrt{\alpha_t}}\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\epsilon_0$, and $\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\hat{\epsilon}_\theta(\mathbf{x}_t, t)$, so the NN learns to predict the added noise.
Training: img $\mathbf{x}_0$, $t \sim \mathrm{Unif}(1 \dots T)$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, GD on $\nabla_\theta \left\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\right\|^2$.
Sampling: $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$, for $t = T$ downto 1: $\mathbf{z} \sim \mathcal{N}(0, I)$ if $t > 1$ else $\mathbf{z} = 0$;
$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)) + \sigma_t \mathbf{z}$.
$\sigma_t^2 = \beta_t$ in practice. t can be continuous.

8.1 Conditional generation

Add input y to the model.
ControlNet: don't retrain model, add layers that add something to block outputs.
(Classifier-free) **guidance**: mix predictions of a conditional and unconditional model, because conditional models are not diverse. $\eta_{\theta_1}(x, c; t) = (1+\rho)\eta_{\theta_1}(x, c; t) - \rho\eta_{\theta_2}(x; t)$.

8.2 Latent diffusion models

High-res images are expensive to model. Predict in latent space, decode with a decoder.

9 Reinforcement learning

Environment is a Markov Decision Process: states S, actions A, reward $r : S \times A \to \mathbb{R}$, transition $p : S \times A \to S$, initial $s_0 \in S$, discount factor γ. r and p are deterministic, can be a distribution. Learn policy $\pi : S \to A$. Value $V_\pi : S \to \mathbb{R}$, the reward from s under π.
Bellman eq.: $G_t := \sum_{k=0}^\infty \gamma^k R_{t+k+1}$, $v_\pi(s) := \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s] = \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a)[r + \gamma \mathbb{E}_\pi[G_{t+1} \mid S_{t+1} = s']] = \sum_a \pi(a \mid s) \sum_{s', r} p(s', r \mid s, a)[r + \gamma v_\pi(s')]$. Can be solved via dynamic programming (needs knowledge of p), Monte-Carlo or Temporal Difference learning.

9.1 Dynamic programming

Value iteration: compute optimal v_*, then π_*.
Policy iteration: compute v_π and π together.
For any V_π the greedy policy (optimal) is $\pi'(s) = \arg\max_{a \in A}(r(s,a) + \gamma V_\gamma(p(s,a)))$.
Bellman optimality: $v_*(s) = \max_a q_*(s,a) = \max_a \sum_{s',r} p(s', r \mid s, a)[r + \gamma v_*(s')] \Rightarrow$ update step: $V^*_{\text{new}}(s) = \max_{a \in A}(r(s,a) + \gamma V^*_{\text{old}}(s'))$, when $V^*_{\text{old}} = V^*_{\text{new}}$, we have optimal policy.
Converges in finite steps, more efficient than policy iteration. But needs knowledge of p, iterates over all states and $O(|S|)$ memory.

9.2 Monte Carlo sampling

Sample trajectories, estimate v_π by averaging returns. Doesn't need full p, is unbiased, but high variance, exploration/exploitation dilemma, may not reach term. state.

9.3 Temporal Difference learning

For each $s \to s'$ by action a update: $\Delta V(s) = r(s,a) + \gamma V(s') - V(s)$. **$\varepsilon$-greedy policy**: with prob. ε choose random action, else greedy.

9.4 Q-learning

Q-value f.: $q_\pi(s,a) := \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$.
SARSA (on-policy): For each $S \to S'$ by action A update: $\Delta Q(S,A) = r(S,A) + \gamma Q(S',A') - Q(S,A)$, $Q(S,A) \mathrel{+}= \alpha \Delta Q(S,A)$, α is LR.
Q-learning (off-policy/offline): $\Delta Q(S,A) = R_{t+1} + \gamma \max_a Q(S',a) - Q(S,A)$
All these approaches do not approximate values of states that have not been visited.

9.5 Deep Q-learning

Use NN to predict Q-values. Loss is $(R + \gamma \max_{a'} Q_\theta(S',a') - Q_\theta(S,A))^2$, backprop only through $Q_\theta(S,A)$. Store history in replay buffer, sample from it for training $\Rightarrow$ no correlation in samples.

9.6 Deep Q-networks

Encode state to low dimensionality with NN.

9.7 Policy gradients

Q-learning does not handle continuous action spaces. Learn a policy directly instead, $\pi(a_t \mid s_t) = \mathcal{N}(\mu_t, \sigma_t^2 \mid s_t)$. Sample trajectories: $p(\tau) = p(s_1, a_1, \ldots, s_T, a_T) = p(s_1) \prod \pi(a_t|s_t) p(s_{t+1} | a_t, s_t)$. This is on-policy.
Eval: $J(\theta) := \mathbb{E}_{\tau \sim p_\theta(\tau)}[\sum_t \gamma^t r(s_t, a_t)]$. To optimize, need to compute $\mathbb{E}$ (see proofs).
REINFORCE: MC sampling of τ. To reduce variance, subtract baseline $b(s_t)$ from reward.

9.8 Actor-Critic

$\nabla_\theta J(\theta) = \frac{1}{N} \sum_i \sum_t \nabla \log \pi_\theta(a_t^i \mid s_t^i)(r(s_t^i, a_t^i) + \gamma V(s_{t+1}^i) - V(s_t^i)))$. π = actor, V = critic. Est. value with NN, not traj. rollouts.

9.9 Motion synthesis

Data-driven: bad perf. out of distribution, needs expensive mocap.
DeepMimic: RL to imitate reference motions while satisfying task objectives.
SFV: use pose estimation: videos $\to$ train data.

10 Neural Implicit Representations

Voxels/volum. primitives are inefficient (n^3 compl.). Meshes have limited granularity and have self-intersections. **Implicit representation**: $S = \{x \mid f(x) = 0\}$. Can be invertibly transformed without accuracy loss. Usually represented as signed distance function values on a grid, but this is again n^3. By UAT, approx. f with NN. **Occupancy networks**: predict probability that point is inside the shape. **DeepSDF**: predict SDF. Both conditioned on input (2D image, class, etc.). Continuous, any topology/resolution, memory-efficient. NFs can model other properties (color, force, etc.).

10.1 Learning 3D Implicit Shapes

Inference: to get a mesh, sample points, predict occupancy/SDF, use marching cubes.

10.1.1 From watertight meshes

Sample points in space, compute GT occupancy/SDF, CE loss.

10.1.2 From point clouds

Only have samples on the surface. Weak supervision: loss $= |f_\theta(x_i)|^2 + \lambda \mathbb{E}_x(\|\nabla_x f_\theta(x)\| - 1)^2$, edge points should have $\|\nabla f\| \approx 1$ by def. of SDF, $f \approx 0$.

10.1.3 From images

Need differentiable rendering 3D $\to$ 2D. **Differentiable Volumetric Rendering**: for a point conditioned on encoded image, predict occupancy $f(x)$ and RGB color $c(x)$. **Forward**: for a pixel, raymarch and root find $\hat{p} : f(\hat{p}) = 0$ with secant. Set pixel color to $c(\hat{p})$. **Backward**: see proofs.

10.2 Neural Radiance Fields (NeRF)

$(x, y, z, \theta, \phi) \xrightarrow{\text{NN}} (r, g, b, \sigma)$. Density is predicted before adding view direction θ, ϕ, then one layer for color. **Forward**: shoot ray, sample points along it and blend: $\alpha := 1 - \exp(-\sigma_i \delta_i)$, $\delta_i := t_{i+1} - t_i$, $T_i := \prod_{j=1}^{i-1}(1 - \alpha_j)$, color is $c = \sum_i T_i \alpha_i c_i$. Optimized on many views of the scene. Can handle transparency/thin structure, but worse geometry. Needs many (50+) views for training, slow rendering for high res, only models static scenes.

10.2.1 Positional Encoding for High Frequency Details

Replace x, y, z with pos. enc. or rand. Fourier feats. Adds high frequency feats.

10.2.2 NeRF from sparse views

Regularize geometry and color.

10.2.3 Fast NeRF render. and train.

Replace deep MLPs with learn. feature hash table + small MLP. For x interp. features between corners.

10.3 3D Gaussian Splatting

Alternative parametr.: Find a cover of object with primitives, predict inside. Or sphere clouds. Both ineff. for thin structures. Ellipsoids are better.
Initialize point cloud randomly or with an approx. reconstruction. Each point has a 3D Gaussian. Use camera params. to project ("splat") Gaussians to 2D and differentiably render them. Adaptive density control moves/clones/merges points.
Rasterization: for each pixel sort Gaussians by depth, opacity $\alpha = o \cdot \exp(-0.5(x - \mu')^\top \Sigma'^{-1}(x - \mu'))$, rest same as NeRF.
Each Gaussian primitive has a center $\mu \in \mathbb{R}^3$, a covariance $\Sigma \in \mathbb{R}^{3\times3}$, and a color $c \in \mathbb{R}^3$ and an opaciti $o \in \mathbb{R}$.
To model view-dependent color, the color can be replaced with spherical harmonics, i.e. $c \in \mathbb{R}^9$
To keep covariance semi-definite: $\Sigma = RSS^\top R^\top$, where $S \in \mathbb{R}^{3\times3}$ is a diagonal scaling matrix and $R \in \mathbb{R}^{3\times3}$ is a rotation matrix.

11 Parametric body models

11.1 Pictorial structure

Unary terms and pairwise terms between them with springs.

11.2 Deep features

Direct regression: predict joint coordinates with refinement.
Heatmaps: predict probability for each pixel, maybe Gaussian. Can do stages sequentially.

11.3 3D

Naive 2D $\to$ 3D lift works. But can't define constraints $\Rightarrow$ 2m arms sometimes.
Skinned Multi-Person Linear model (SMPL) is the standard non-commerical model. 3D mesh, base mesh is ~7k vertices, designed by an artist. To produce the model, point clouds (scans) need to be aligned with the mesh. **Shape deformation subspace**: for a set of human meshes T-posing, vectorize their vertices T and subtract the mean mesh. With PCA represent any person as weighted sum of 10-300 basis people, $T = S\beta + \mu$.
For pose, use **Linear Blend Skinning**. $\mathbf{t}_i' = \sum_k w_{ki} \mathbf{G}_k(\boldsymbol{\theta}, \mathbf{J}) \mathbf{t}_i$, where $\mathbf{t}$ is the T-pose positions of vertices, $\mathbf{t}'$ is transformed, w are weights, $\mathbf{G}_k$ is rigid bone transf., $\boldsymbol{\theta}$ is pose, $\mathbf{J}$ are joint positions. Linear assumption produces artifacts. **SMPL**: $\mathbf{t}_i' = \sum_k w_{ki} \mathbf{G}_k(\boldsymbol{\theta}, \mathbf{J}(\boldsymbol{\beta}))(\mathbf{t}_i + \mathbf{s}_i(\boldsymbol{\beta}) + \mathbf{p}_i(\boldsymbol{\theta}))$. Adds shape correctives $\mathbf{s}(\boldsymbol{\beta}) = \mathbf{S}\boldsymbol{\beta}$, pose cor. $\mathbf{p}(\boldsymbol{\theta}) = \mathbf{P}\boldsymbol{\theta}$, $\mathbf{J}$ dep. on shape $\boldsymbol{\beta}$.
Predicting human pose is just predicting $\boldsymbol{\beta}, \boldsymbol{\theta}$ and camera parameters.

11.3.1 Optimization-based fitting

Predict 2D joint locations, fit SMPL to them by argmin with prior regularization. Argmin is hard to find, learn F: $\Delta\theta = F(\frac{\partial L_{reproj}}{\partial \theta}, \theta^t, x)$. Issues: self-occlusion, no depth info, non-rigid deformation (clothes).

11.3.2 Template-based capture

Scan for first frame, then track with SMPL.

11.3.3 Animatable Neural Implicit Surfaces

Model base shape and w with 2 NISs.

12 ML

Perceptron converges in finite time iff data is linearly separable. **MAP** $\theta^* \in \arg\max p(\theta \mid X, y)$. **MLE** $\theta \in \arg\max p(y \mid X, \theta)$ consistent, efficient. **Binary cross-entropy** $L(\theta) = -y_i \log(\hat{y}_i) - (1-y_i)\log(1 - \hat{y}_i)$. Cross-entropy $H(p_d, p_m) = H(p_d) + D_{\mathrm{KL}}(p_d \| p_m)$.
For any continuous f $\exists$NN $g(x)$, $|g(x) - f(x)| < \varepsilon$. 1 hidden layer is enough, activation function needs to be nonlinear.
MLP backward inputs: $\delta^{(l)} = \delta^{(l+1)} \cdot \frac{\partial \mathbf{z}^{(l+1)}}{\partial \mathbf{z}^{(l)}}$, backward weights: $\left[\frac{\partial \mathbf{z}^{(l)}}{\partial \mathbf{w}_{ij}^{(l)}}\right]_k = f'(\mathbf{a})_k \cdot z_j^{(l)} \cdot [k = i]$, $\frac{\partial^2 L}{\partial \mathbf{z}^{(l)} \partial \mathbf{z}^{(l)}} = \delta^{(l)} \frac{\partial \mathbf{z}^{(l)}}{\partial \mathbf{w}_{i,j}^{(l)}}$, backward bias: $\frac{\partial L}{\partial \mathbf{b}_i^{(l)}}$ = same, but no $\mathbf{z}$.

12.1 Activation functions

name	$f(x)$	$f'(x)$	$f(X)$
sigmoid	$\frac{1}{1+e^{-x}}$	$\sigma(x)(1-\sigma(x))$	$(0,1)$
tanh	$\frac{e^x - e^{-x}}{e^x + e^{-x}}$	$1 - \tanh(x)^2$	$(-1,1)$
ReLU	$\max(0,x)$	$[x \geq 0]$	$[0,\infty)$

Finite range: stable training, mapping to prob. space. Sigmoid, tanh saturate (value with large mod have small gradient) $\Rightarrow$ vanishing gradient, Tanh is linear around 0 (easy learn), ReLU can blow up activation; piecewise linear $\Rightarrow$ faster convergence.

12.2 GD algos

SGD: use 1 sample. For sum structured loss is unbiased. High variance, efficient, jumps a lot $\Rightarrow$ may get out of local min., may overshoot. **Mini-batch**: use $m < n$ samples. More stable, parallelized. **Polyak's momentum**: velocity $\mathbf{v} := \alpha \mathbf{v} - \epsilon \nabla_\theta L(\theta)$, $\theta := \theta + \mathbf{v}$. Move faster when high curv., consistent or noisy grad. **Nesterov's momentum**: $\mathbf{v} := \alpha \mathbf{v} - \epsilon \nabla_\theta L(\theta + \alpha \mathbf{v})$. Gets grad. at future point. **AdaGrad**: $\mathbf{r} := \mathbf{r} + \nabla \odot \nabla$, $\Delta\theta = -\epsilon / (\delta + \sqrt{\mathbf{r}}) \odot \nabla$. Grads decrease fast for variables with high historical gradients, slow for low. But can decrease LR too early/-fast. **RMSProp**: $\mathbf{r} := \rho\mathbf{r} + (1-\rho)\nabla\odot\nabla$, use weighted moving average ⇒ drop history from distant past, works better for nonconvex. **Adam**: collect 1st and 2nd moments: $\mathbf{m} := \beta_1\mathbf{m} + (1-\beta_1)\nabla, \mathbf{v} := \beta_2\mathbf{v}+(1-\beta_2)\nabla\odot\nabla$, unbias: $\hat{\mathbf{m}} = \mathbf{m}/(1-\beta_1^t), \hat{\mathbf{v}} = \mathbf{v}/(1-\beta_2^t)$, $\Delta\theta = -\frac{\eta}{\sqrt{\hat{\mathbf{v}}+\epsilon}}\hat{\mathbf{m}}$.

13 Proofs

Softmax derivative Let $\hat y_i = f(x)_i = \frac{\exp(x_i)}{\sum_{j=1}^d \exp(x_j)}, x\in\mathbb{R}^d$, y is 1-hot $\in \mathbb{R}^d$, negative log-likelihood $L(\hat y, y) = -\sum_{i=1}^d y_i\log\hat y_i$. $\frac{\partial L}{\partial x_i} = \frac{\partial L}{\partial \hat y}\frac{\partial \hat y}{\partial x_i}$. | $\frac{\partial \hat y_i}{\partial x_i} = \frac{\exp(x_i)\sum_{j=1}^d\exp(x_j) - \exp^2(x_i)}{\left(\sum_{j=1}^d \exp(x_j)\right)^2} = \frac{\exp(x_i)}{\sum_{j=1}^d\exp(x_j)}\left(\frac{\sum_{j=1}^d\exp(x_j)}{\sum_{j=1}^d \exp(x_j)} - \frac{\exp(x_i)}{\sum_{j=1}^d\exp(x_j)}\right) = \hat y_i(1-\hat y_i)$. | $\frac{\partial \hat y_k}{\partial x_i} = \frac{-\exp(x_i)\exp(x_k)}{\left(\sum_{j=1}^d\exp(x_j)\right)^2} = -\hat y_i\hat y_k$. | $\frac{\partial L}{\partial \hat y_k} = -\frac{y_k}{\hat y_k}$. | $\frac{\partial L}{\partial x_i} = -\frac{y_i}{\hat y_i}(\hat y_i(1-\hat y_i)) - \sum_{k\neq i}\frac{y_k}{\hat y_k}(-\hat y_i\hat y_k) = -y_i + y_i\hat y_i + \sum_{k\ne i} y_k\hat y_i = -y_i + \hat y_i\sum_k y_k = \hat y_i - y_i$.

BPTT ρ is the identity function, ∂^+ is the immediate derivative, ignoring the effect from recurrence. $\frac{\partial \mathbf{h}_t}{\partial W} = \frac{\partial}{\partial W}\mathbf{h}_t(\rho(W), \mathbf{h}_{t-1}(W)) = \frac{\partial \mathbf{h}_t}{\partial\rho}\frac{\partial\rho}{\partial W} + \frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_{t-1}}\frac{\partial \mathbf{h}_{t-1}}{\partial W} = \frac{\partial^+\mathbf{h}_t}{\partial W} + \frac{\partial\mathbf{h}_t}{\partial\mathbf{h}_{t-1}}\frac{\partial\mathbf{h}_{t-1}}{\partial W} = \frac{\partial^+\mathbf{h}_t}{\partial W} + \frac{\partial \mathbf{h}_t}{\partial\mathbf{h}_{t-1}}\left[\frac{\partial^+\mathbf{h}_{t-1}}{\partial W} + \frac{\partial \mathbf{h}_{t-1}}{\partial\mathbf{h}_{t-2}}\frac{\partial^+\mathbf{h}_{t-2}}{\partial W}+\dots\right] = \sum_{k=1}^t \frac{\partial\mathbf{h}_t}{\partial\mathbf{h}_k}\frac{\partial^+\mathbf{h}_k}{\partial W}$ | $\frac{\partial L_t}{\partial W} = \frac{\partial L_t}{\partial \hat y_t}\frac{\partial\hat y_t}{\partial \mathbf{h}_t}\sum_{k=1}^t\frac{\partial\mathbf{h}_t}{\partial\mathbf{h}_k}\frac{\partial^+\mathbf{h}_k}{\partial W}$

BPTT divergence Let λ_1 be the largest singular value of $\mathbf{W}_{hh}$, $\|\mathrm{diag} f'(\mathbf{h}_{i-1})\| < \gamma, \gamma\in\mathbb{R}$, $\|\cdot\|$ is the spectral norm. If $\lambda_1 < \gamma^{-1}$, then $\forall i \left\|\frac{\partial \mathbf{h}_i}{\partial\mathbf{h}_{i-1}}\right\| \le \|\mathbf{W}_{hh}^\top\|\|\mathrm{diag} f'(\mathbf{h}_{i-1})\| < \frac{1}{\gamma}\gamma < 1 \Rightarrow \exists\eta : \forall i \left\|\frac{\partial\mathbf{h}_i}{\partial\mathbf{h}_{i-1}}\right\| \le \eta < 1$, by induction over i: $\left\|\prod_{i=k+1}^t \frac{\partial\mathbf{h}_i}{\partial\mathbf{h}_{i-1}}\right\| \le \eta^{t-k}$, so the gradients vanish as $t\to\infty$. Similarly, if $\lambda_1>\gamma^{-1}$, then gradients explode.

$D_{\mathrm{KL}}(\cdot\|\cdot)\ge 0$ $-D_{\mathrm{KL}}(p\|q) = -\mathbb{E}_{x\sim p}\log\frac{p(x)}{q(x)} = \mathbb{E}\log\frac{q(x)}{p(x)} \le \log\mathbb{E}_{x\sim p}\frac{q(x)}{p(x)} = \log\int q(x)\mathrm{d}x = \log 1 = 0$.

VAE ELBO $\log p_\theta(x^{(i)}) = \mathbb{E}_{z\sim q_\phi(z|x^{(i)})}\log p_\theta(x^{(i)}) = \mathbb{E}_z\log\frac{p_\theta(x^{(i)}|z)p_\theta(z)}{p_\theta(z|x^{(i)})} = \mathbb{E}_z\log\frac{p_\theta(x^{(i)}|z)p_\theta(z)q_\phi(z|x^{(i)})}{p_\theta(z|x^{(i)})q_\phi(z|x^{(i)})} = \mathbb{E}_z\log p_\theta(x^{(i)}\mid z) - \mathbb{E}_z\log\frac{q_\phi(z|x^{(i)})}{p_\theta(z)} + \mathbb{E}_z\log\frac{q_\phi(z|x^{(i)})}{p_\theta(z|x^{(i)})} = \mathbb{E}_z\log p_\theta(x^{(i)}\mid z) - D_{\mathrm{KL}}(q_\phi(z\mid x^{(i)})\|p_\theta(z)) + D_{\mathrm{KL}}(q_\phi(z\mid x^{(i)})\|p_\theta(z\mid x^{(i)}))$.

KL for ELBO Let $p(z) = \mathcal{N}(0,\mathbf{I}), q(z\mid x) = \mathcal{N}(\mu,\sigma^2\mathbf{I}), J := \dim z$. By $\int p(z)\log q(z)\mathrm{d}z = -\frac{J}{2}\log 2\pi - \frac12\sum_{j=1}^J\log\sigma_{q,j}^2 - \frac12\sum_{j=1}^J\frac{\sigma_{p,j}^2+(\mu_{p,j}-\mu_{q,j})^2}{\sigma_{q,j}^2}$ we have $\int q(z\mid x)\log p(z)\mathrm{d}z = -\frac{J}{2}\log 2\pi - \frac12\sum_{j=1}^J(\sigma_j^2+\mu_j^2)$ and $\int q(z\mid x)\log q(z\mid x)\mathrm{d}z = -\frac{J}{2}\log 2\pi - \frac12\sum_{j=1}^J(\log\sigma_j^2+1)$, so $-D_{\mathrm{KL}}(q(z\mid x)\|p(z)) = \frac12\sum_{j=1}^J(1+\log\sigma_j^2-\mu_j^2-\sigma_j^2)$

Optimal discriminator D^* maximizes $V(G,D) = \int_x p_d\log D(x)\mathrm{d}x + \int_z p(z)\log(1-D(G(Z)))\mathrm{d}z = \int_x p_d\log D(x)\mathrm{d}x + p_m(x)\log(1-D(x))\mathrm{d}z$, and for $f(y) = a\log(y) + b\log(1-y)$: $f'(y) = \frac{a}{y} - \frac{b}{1-y} \Rightarrow f'(y)=0 \Leftrightarrow y = \frac{a}{a+b}$, $f''(\frac{a}{a+b}) = -\frac{a}{(\frac{a}{a+b})^2} - \frac{b}{(1-\frac{a}{a+b})^2} < 0$ for $a,b>0 \Rightarrow$ max. at $\frac{a}{a+b} \Rightarrow D^* = \frac{p_d(x)}{p_d(x)+p_m(x)}$

Expectation of reparam. $\nabla_\varphi\mathbb{E}_{p_\varphi(z)}(f(z)) = \nabla_\varphi\int p_\varphi(z)f(z)\mathrm{d}z = \nabla_\varphi\int p_\varphi(z)f(z)\mathrm{d}z = \nabla_\varphi\int p_\varphi(z)f(g(\epsilon,\varphi))\mathrm{d}\epsilon = \mathbb{E}_{p(\epsilon)}\nabla_\varphi f(g(\epsilon,\varphi))$

$q(x_t\mid x_0)$ $\mathbf{x}_t = \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\epsilon = \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\epsilon = \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\epsilon = \dots = \sqrt{\bar\alpha_t}\mathbf{x}_0 + \sqrt{1-\bar\alpha_t}\epsilon$

Bellman operator converges Want to prove that value iteration converges to the optimal policy: $\lim_{k\to\infty}(T^*)^k(V) = V_*$, where $T^*(V) = \max_{a\in A}\sum_{s',r}p(s',r\mid s,a)(r(s,a)+\gamma V(s'))$. T^* is a contraction mapping, i.e. $\max_{s\in S}|T^*(V_1(s)) - T^*(V_2(s))| \le \gamma\max_{s\in S}|V_1(s)-V_2(s)|$: LHS $\le \max_{s,a}|\sum_{s',r}p(s',r\mid s,a)(r(s,a)+\gamma V_1(s')) - \sum_{s',r}p(s',r\mid s,a)(r(s,a)+\gamma V_2(s'))| = \gamma\max_{s,a}|\sum_{s',r}p(s',r\mid s,a)(V_1(s')-V_2(s'))| =$ RHS. By the contraction th., T^* has a unique fixed point, and we know V^* is a FP of T^*. As $\gamma<1$, LHS$(V,V^*)\to 0$ and $T^*(V)\to V_*$.

Policy gradients $J(\theta) = \mathbb{E}_{\tau\sim p(\tau)}[r(\tau)] = \int p(\tau)r(\tau)\mathrm{d}\tau$. $\nabla_\theta J(\theta) = \int\nabla_\theta p(\tau)r(\tau)\mathrm{d}\tau = \int p(\tau)\nabla_\theta\log p(\tau)r(\tau)\mathrm{d}\tau = \mathbb{E}_{\tau\sim p(\tau)}[\nabla_\theta\log p(\tau)r(\tau)] = \mathbb{E}_{\tau\sim p(\tau)}[\nabla_\theta\log p(\tau)r(\tau)]$. $\log p(\tau) = \log[p(s_1)\prod\pi_\theta(a_t\mid s_t)p(s_{t+1}\mid a_t,s_t)] = 0 + \sum_t\log\pi_\theta(a_t\mid s_t) + 0$
$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim p(\tau)}[(\sum_t\nabla\log p_\theta(a_t^i\mid s_t^i))(\sum_t\gamma^t r(s_t^i,a_t^i))]$. max likelihood, trajectory reward scales the gradient.

Implicit differentiation $\frac{\mathrm{d}y}{\mathrm{d}x}$ of $x^2+y^2=1$: $\frac{\mathrm{d}}{\mathrm{d}x}(x^2+y^2) = \frac{\mathrm{d}}{\mathrm{d}x}(1) \Rightarrow \frac{\mathrm{d}}{\mathrm{d}x}x^2 + \frac{\mathrm{d}}{\mathrm{d}x}y^2 = 0 \Rightarrow 2x + (\frac{\mathrm{d}}{\mathrm{d}y}y^2)\frac{\mathrm{d}y}{\mathrm{d}x} = 0\Rightarrow 2x+2y\frac{\mathrm{d}y}{\mathrm{d}x}=0\Rightarrow\frac{\mathrm{d}y}{\mathrm{d}x} = -\frac{x}{y}$

DVR Backward pass $\frac{\partial L}{\partial\theta} = \sum_u\frac{\partial L}{\partial\hat{\mathbf{I}}_u}\cdot\frac{\partial\hat{\mathbf{I}}_u}{\partial\theta}$ | $\frac{\partial\hat{\mathbf{I}}_u}{\partial\theta} = \frac{\partial c_\theta(\hat{\mathbf{p}})}{\partial\theta} + \frac{\partial t_\theta(\hat{\mathbf p})}{\partial\hat{\mathbf p}}\cdot\frac{\partial\hat{\mathbf p}}{\partial\theta}$. Ray $\hat{\mathbf p} = r_0 + \hat d\mathbf{w}$, r_0 is camera pos., $\mathbf w$ is ray dir., $\hat d$ is ray dist. Implicit def.: $f_\theta(\hat{\mathbf p}) = \tau$. Diff.: $\frac{\partial f_\theta(\hat{\mathbf p})}{\partial\theta} + \frac{\partial f_\theta(\hat{\mathbf p})}{\partial\hat{\mathbf p}}\cdot\frac{\partial\hat{\mathbf p}}{\partial\theta} = 0 \Rightarrow \frac{\partial f_\theta(\hat{\mathbf p})}{\partial\theta} + \frac{\partial f_\theta(\hat{\mathbf p})}{\partial\hat{\mathbf p}}\cdot\mathbf{w}\frac{\partial\hat d}{\partial\theta} = 0\Rightarrow\frac{\partial\hat{\mathbf p}}{\partial\theta} = \mathbf w\frac{\partial\hat d}{\partial\theta} = -\mathbf w(\frac{\partial f_\theta(\hat{\mathbf p})}{\partial\hat{\mathbf p}}\cdot\mathbf w)^{-1}\frac{\partial f_\theta(\hat{\mathbf p})}{\partial\theta}$

14 Appendix

Secant Method Line $(x_0,f(x_0))\to(x_1,f(x_1))$, approx.: $y = \frac{f(x_1)-f(x_0)}{x_1-x_0}(x-x_1)+f(x_1)$, $y=0$ at $x_2 = x_1 - f(x_1)\frac{x_1-x_0}{f(x_1)-f(x_0)}$. Approximates Newton's method without derivatives.

Implicit plane from 3 points $(x_1,0,0),(0,y_1,0),(0,0,z_1)\Rightarrow x/x_1+y/y_1+z/z_1-1=0$. More generally: let a,b any vectors on plane, $n := a\times b = (a_2b_3-a_3b_2, a_3b_1-a_1b_3, a_1b_2-a_2b_1)\Rightarrow n_1x+n_2y+n_3z+k=0$, subst. any point to find k.

Torus equation $(\sqrt{x^2+y^2}-R)^2+z^2=r^2$, cent. 0, around z axis.

Chain rule $\frac{\partial z_k}{\partial x_j} = \sum_{i=1}^m\frac{\partial z_k}{\partial y_i}\frac{\partial y_i}{\partial x_j}$ | $\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y}\frac{\partial y}{\partial x}$

Derivatives $(f\cdot g)' = f'g+fg'$, $(f/g)' = (f'g-fg')/g^2$, $(f\circ g)' = f'(g)g'$, $(f^{-1})' = 1/f'(f^{-1})$, $(\log x)' = 1/x$.

Linear algebra $\det(\mathbf A+\mathbf{u}\mathbf{v}^\top) = (1+\mathbf v^\top\mathbf A^{-1}\mathbf u)\det\mathbf A$

Jensen's inequality $f(\mathbb E[X])\le\mathbb E[f(X)]$ and $f(tx_1+(1-t)x_2)\le tf(x_1)+(1-t)f(x_2)$ if f is convex, i.e. $\forall t\in[0,1], x_1,x_2\in X$: $f(tx_1+(1-t)x_2)\le tf(x_1)+(1-t)f(x_2)$.

Gaussians $\mathcal N(\mu_1,\Sigma_1)+\mathcal N(\mu_2,\Sigma_2) = \mathcal N(\mu_1+\mu_2,\Sigma_1+\Sigma_2)$, $a\cdot\mathcal N(\mu,\Sigma) = \mathcal N(a\mu,a^2\Sigma)$. $\mathcal N = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp(-\frac12(x-\mu)^\top\Sigma^{-1}(x-\mu))$

VRNN $p_\theta(\mathbf z) = \prod_{t=1}^T p_\theta(z_t\mid\mathbf z_{<t},\mathbf x_{<t})$, $q_\phi(\mathbf x\mid\mathbf z) = \prod_{t=1}^T q_\phi(z_t\mid\mathbf x_{\le t},\mathbf z_{<t})$, $p_\theta(\mathbf x,\mathbf z) = \prod_{t=1}^T p_\theta(x_t\mid\mathbf z_{\le t},\mathbf x_{<t})p_\theta(z_t\mid\mathbf z_{<t},\mathbf x_{<t})$.

Misc A **translation vector** is added. **Bayes rule**: $P(A\mid B) = P(B\mid A)P(A)/P(B)$. A function f is **volume preserving** if $\left|\det\frac{\partial f^{-1}(x)}{\partial x}\right| = 1$. **Negative log-likelihood** $L(\hat y,y) = -\sum_i y_i\log\hat y_i$