Plan:

1. Vector Q. basics
   - → Hamming game
   - → MXFP4
   - → Lloyd-Max

2. Modern Approaches:
   - → VQ-VAE, VQ-GAN
   - → Gumbel-Softmax (DALL-E1, ...)
   - → FSQ
   - → RQ-VAE. (audio)

3. Info-theoretic quant.
   - → Lower bound for iid
   - → Upper bound statement.

1a

Last time: $x \in \mathbb{R} \to \hat{x}$ - discrete

Now: $X \in \mathbb{R}^n \to \widehat{X} \in \mathbb{R}^n$ - discrete.

Why?

Go to LDLQ.

(1b) But even when loss is MSE and data is iid there are surprises.

Last time we found out that if $X \in [-1/2; 1/2]$

- Uniform quant at rate R (with dithering) gives $D_q = \frac{2^{-2R}}{12}$.

- Non-Unif quant. gives $D_{nq} \approx \frac{2^{-2R}}{12} \left(\int p_X^{1/3} \right)^3$

(3) To everyone's shock Shannon:

$$(X_1, \dots, X_n) \longrightarrow [2^{nR}]_{\text{pts}}$$

can do better even if $X_j \sim$ iid.

(a) In part. for $X \sim \mathcal{N}(0, \sigma_x^2)$

$D_{\text{vector}}(R) \approx \sigma^2 \cdot 2^{-2R}$

4.dB better!

(a) Hamming game 100 → 50 25%

11%!

why? • $\square$ vs $\bigcirc$ for D_j

• Conc. of meas. — iid kills tails

• Simple demo: worse than

(b) ...

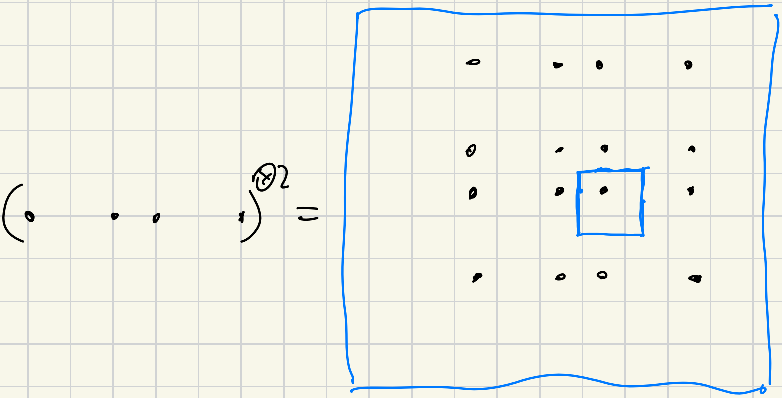
- The shocking thing is that you can do a lot better by jointly quantizing n entries together

- Why? More elbow room: 2^{nR} pts in $\mathbb{R}^n$! (not restricted to rectangles)

- Simple way to use this freedom is in MXFP4: $(X_1, \dots, X_{32}) = 2^{-127 \dots 127} \times (FP4, \dots, FP4)$

Note: 2^{128} scaling $\Rightarrow$ NaN!

$\Rightarrow$ actual rate is $R = 4.25$ bit/par



product constellation

- MXFP4 is a union of dyadically scaled product const. This is good for H/W.

- Let us think more conceptually

- Quantization function: $q: \mathbb{R}^n \to \mathbb{R}^n$

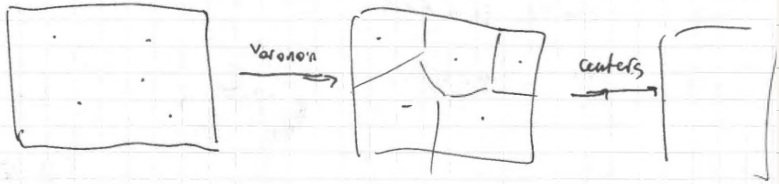
$$q(x) = \sum_{j=1}^{2^{nR}} c_j \mathbb{1}_{D_j}$$

Lloyd-Max algo:

1) Given $\{c_j\}$ best D_j = Voronoi region.

2) Given D_j, best $c_j = \mathbb{E}[x \mid x \in D_j]$ - Centroid

Under MSE distortion: given



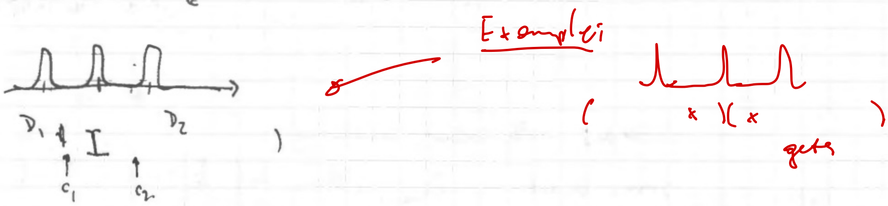
Iterate → Lloyd's algo.

$\boxed{\text{Th}}$ 1) L.a. converges to CVT

2) opt'l quantizer is a CVT

3) ∃ CVT non-optimal.

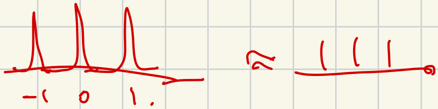
Example:



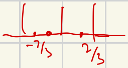
Rmk: Finding opt'l quantizer is not fun

e.g. fun for $N > 3$, $P_X = \mathcal{N}(0, \sigma^2)$

Consider $P_X =$

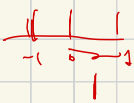


Symmetric: $C_\pm = \pm 2/3$



$(1/3)^2 \, 2/3 + 1/3 \, (2/3)^2 = \frac{2}{9} \cdot \left(\frac{1}{3} + 2/3\right) = 2/9 > 1/6$

Asymmetric: $C_- = -1$, $C_+ = 1/2$



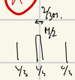
$(1/2)^2 \cdot 2/3 = 2/12 = 1/6$

≡ Better!

Note: this example is not great because suboptimal CUT is „unstable"

TODO: check. (A) **checked!**

Checking:



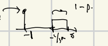
Suppose we set

$D_- = (-\infty, \tau) \Rightarrow$ posterior -1 w.p. $1/3$

w.p. $1/6 + M\tau/2$

$1-p$

$$p = \frac{1/3}{1/3 + 1/6 + M\tau/2}$$



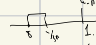
$$C_- = -p + \left(\tau - \tfrac{1}{3M}\right)\tfrac{1}{2}\cdot(1-p) = \left(\tau - \tfrac{1}{3M}\right)\tfrac{1}{2} - p\cdot\left(1 + \tau - \tfrac{1}{3M}\right)\tfrac{1}{2}$$

$D_+ = (\tau, +\infty)$ posterior.

$+1$ w.p. $1/3$

w.p. $\left(\tfrac{1}{3M} - \tau\right)\cdot M/2$.

$$\tilde{p} = \frac{1/3}{1/6 - M\tau/2 + 1/3}$$



$$C_+ = \tilde{p} + (1-\tilde{p})\left(\frac{\tau + 1/3M}{2}\right)$$

$$\tau' = \frac{C_+ + C_-}{2} = \frac{1}{2}\left(-p + \left(\tau - \tfrac{1}{3M}\right)\tfrac{1}{2}(1-p) + \tilde{p} + \frac{(1-\tilde{p})}{2}\left(\tau + \tfrac{1}{3M}\right)\right)$$

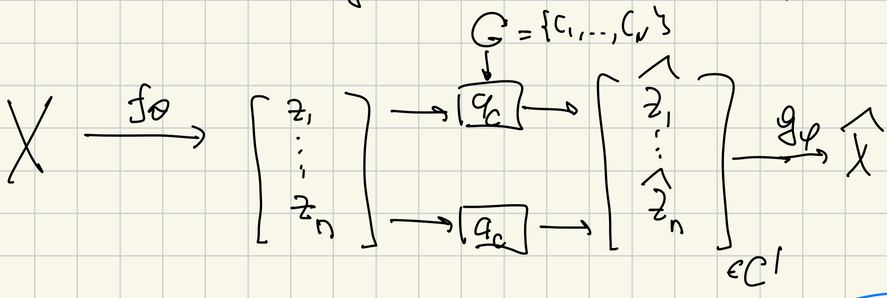
Numerical check: $\left.\frac{d\tau'}{d\tau}\right|_{\tau=0} \gg 1$

$\Rightarrow$ **Unstable!!**

But there must exist an example from Gsn mixtures in $d=3$ and EM getting stuck.

Q: Find such example!

(2) Modern Lloyd-Max VQVAE, VQGAN ....



Loss:

$$L(\theta, \varphi, C) = \hat{\mathbb{E}}\,\|x - \hat{x}\|_2^2$$

Note: f_θ and g_φ operate on many latents, q_c on a single!

does not work gradients do not flow to f_θ!

VQVAE „loss"

$$\|x - g_\varphi(\{\hat{z}\}_{STE})\|^2 + \frac{1}{2}\|\{f_\theta(x)\}_{sg} - \hat{z}\|^2 +$$
$$+ \frac{\beta}{2}\|f_\theta(x) - \{\hat{z}\}_{sg}\|^2$$

- First term:
  - $\nabla_\varphi = \nabla_\varphi \|x - g_\varphi(\hat{z})\|^2$
  - $\nabla_\theta = \sum_j \frac{\partial}{\partial z_j}\Big|_{z=\hat{z}} \|x - g_\varphi(z)\| \cdot \nabla_\theta f_j(x)$
  - $\nabla_C = 0$
- Second term:

  $\nabla_\varphi = \nabla_\theta = 0.$

  $\nabla_{C_j} = \begin{cases} 0, & C_i \neq \hat{z} \\ z, & C_i = \hat{z} \end{cases}$
- Third term: „commitment loss" (forces f_θ to output stuff close to C)

  $\nabla_\varphi = 0$

  $\nabla_C = 0$

  $\nabla_\theta = \frac{\beta}{2}\nabla_\theta \| f_\theta(x)$

VQVAE / VQGAN

- $d = 256 \quad N = 8192. \quad x \sim \mathcal{N}(0, I_n)$

$R = 13/256 \Rightarrow$ Optimal $D = 2^{-2R} \approx 0.93$

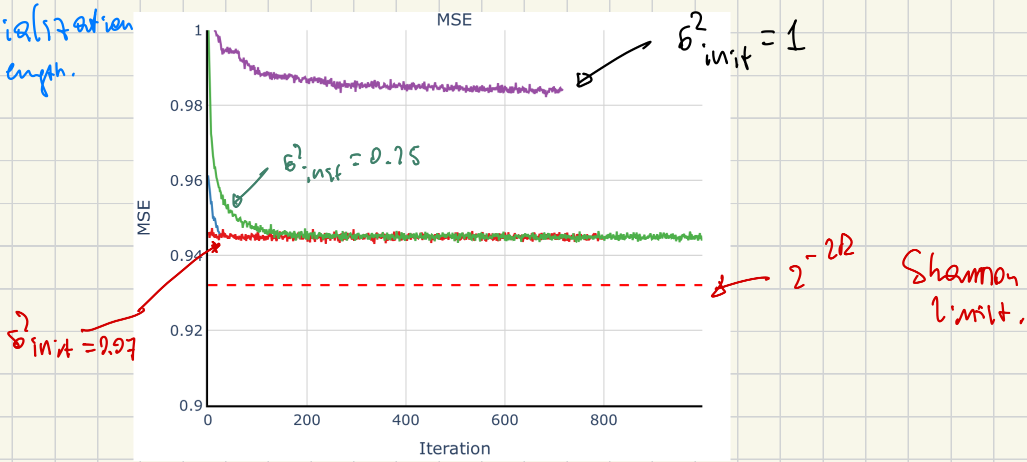
After 60000 of 2000 sized batches

only get 0.984 loss

$\boxed{\text{Note: } \sigma^2_{emb} = 1 - 0.93 = 0.07}$

Note: VQGAN used 1.5M images
(1 img → 16x16 = 256 latents) for 13 epochs
⇒ 1.5B samples to quantize.

Effect of initialization strength.



VQ-ViT

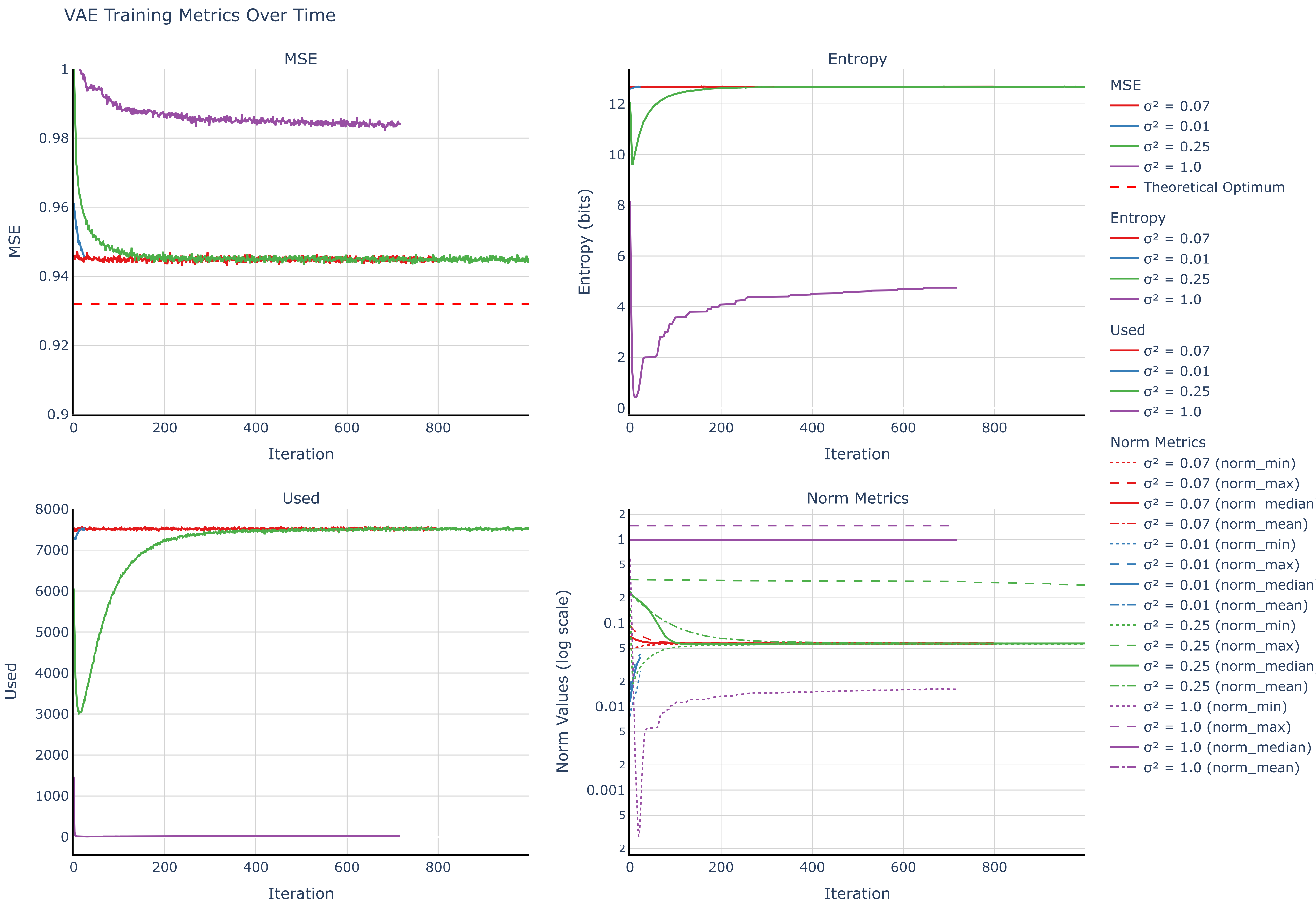
- $d = 32, \quad N = 8192$

$R = 13/32 \Rightarrow D = 2^{-2R} = 0.57.$

Set $B = 16384$ # batches 30000 (0.5B examples)

loss = 0.63

"Better codebook utilization"



VAE Training Metrics Over Time
MSE
Entropy
Used
Norm Metrics
Iteration
Entropy (bits)
Norm Values (log scale)
MSE
σ² = 0.07
σ² = 0.01
σ² = 0.25
σ² = 1.0
Theoretical Optimum
Entropy
Used
Norm Metrics
σ² = 0.07 (norm_min)
σ² = 0.07 (norm_max)
σ² = 0.07 (norm_median)
σ² = 0.07 (norm_mean)
σ² = 0.01 (norm_min)
σ² = 0.01 (norm_max)
σ² = 0.01 (norm_median)
σ² = 0.01 (norm_mean)
σ² = 0.25 (norm_min)
σ² = 0.25 (norm_max)
σ² = 0.25 (norm_median)
σ² = 0.25 (norm_mean)
σ² = 1.0 (norm_min)
σ² = 1.0 (norm_max)
σ² = 1.0 (norm_median)
σ² = 1.0 (norm_mean)

• Let us understand it...

• Study G.D. with ~~STE~~ (STE not needed: no encoder!)

$\{c_1, \dots, c_M\}$

Batch loss $\frac{1}{2} \hat{\mathbb{E}} \|x - q_C(x)\|^2 = \sum_j \hat{\mathbb{E}} \{ \|x - c_j\|^2 \, 1\{x \in D_j\} \}$

$$\nabla_{c_j} = \hat{\mathbb{E}} \, (c_j - x) \, 1\{x \in D_j\}$$

$$= \hat{p}_j \, (c_j - \hat{\mu}_j)$$

empirical   centroid.

$$c_j \leftarrow c_j (1 - \hat{p}_j \eta) + \eta \hat{p}_j \hat{\mu}_j$$

Population step-size modulated by p_j $\Rightarrow$ Small utilization of codebook!

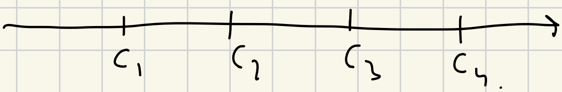
Open q: what if we drop $\hat{p}_j$?

A: [Botou-Bengio '94] Show this corresponds to Newton step in k-means!

Batch size = 1: $X \longrightarrow c_{j^*}$

$$c_{j^*} \leftarrow c_{j^*} (1 - \eta) + \eta X.$$

So given partition:



Sample X $\rightarrow$ update nearest ab towards x

③ Information Theory: Lower Bound

- So let us consider $\underline{X} = (X_1, \dots, X_n) \overset{iid}{\sim} P_X$.
- Suppose there is a quantizer $q(\underline{X})$ with support of size 2^{nR}.

$\boxed{\text{Prop}}$ There exists a coupling $P_{X,Y} \in \mathcal{S}(\mathbb{R} \times \mathbb{R})$ s.t.

$$I(X;Y) \leq R$$

$$\mathbb{E}(X-Y)^2 \leq \frac{1}{n} \mathbb{E} \|\underline{X} - q(\underline{X})\|^2$$

<u>Pf:</u> $\underline{X} = X^n$.

$$I(X^n;U) = H(X^n) - H(X^n|U) = \sum_i \left(H(X_i) - H(X_i|X^{i-1},U)\right) \geq \sum_i H(X_i) - H(X_i|U) = \sum_i I(X_i;U)$$

$$R \geq \frac{1}{n} I(X^n;\hat{X}^n) \geq \frac{1}{n} \sum_i I(X_i;\hat{X}^n) \geq \frac{1}{n} \sum_i I(X_i;\hat{X}_i)$$

Let $\tau \sim \mathrm{Unif}([n]) \perp\!\!\!\perp (X^n, \hat{X}^n)$

$$X = X_\tau$$

$$Y = \hat{X}_\tau$$

$$\frac{1}{n} \sum_i I(X_i;\hat{X}_i) = I(X;Y|\tau) = I(X;Y,\tau) \geq I(X;\tau)$$

$$\mathbb{E}(X-Y)^2 = \mathbb{E} \frac{1}{n} \sum_i (X_i - \hat{X}_i)^2 = \frac{1}{n} \mathbb{E} \|\underline{X} - q(\underline{X})\|^2$$

$\boxed{\text{Corollary}}$ Define

$$D_{vec}(R) \triangleq \inf\{\mathbb{E}(X-Y)^2 : I(X;Y) \le R\}$$

Then for any rate R quantizer

$$\mathbb{E}\|X - q(X)\|^2 \ge n D_{vec}(R)$$

- The function $D_{vec}(R)$ is C^∞, cvx, decr

- Its convex conjugate is "EOT projection"

$$\min_{P_{Y|X}} \mathbb{E}(X-Y)^2 + \lambda I(X;Y)$$

Nice cvx problem!

$$\| $$

$$\min_{P_Y} \min_{P_{X,Y}: P_X \leftrightarrow P_Y} \mathbb{E}(X-Y)^2 + \lambda D(P_{X,Y} \| P_X \times P_Y)$$

$$\| \Delta$$

$$\min_{P_Y} \underbrace{W_2^{(\lambda)}(P_X, P_Y)}_{\text{EOT distance.}}$$

(Sinkhorn algo, aka Blahut-Arimoto)

<u>Example:</u> $P_X = \mathcal{N}(0,1)$.

$$\min \ \mathbb{E}(X-Y)^2 + \lambda I(X;Y).$$

<u>Note:</u> IT fact $\forall\, (X_G, Y)$ let (X_G, Y_G) be a Gaussian version $\left(\text{matching } \mathbb{E}\begin{bmatrix} X \\ Y \end{bmatrix} \text{ and } \operatorname{Cov}\begin{bmatrix} X \\ Y \end{bmatrix}\right)$

Then $I(X;Y) \geq I(X_G;Y_G)$.

Pf

$$I(X;Y) = \sup_{Q_{X|Y}} \mathbb{E} \log \frac{dQ_{X|Y}}{dP_X}$$

$$\geq \mathbb{E}_{X,Y} \log \underbrace{\frac{P_{X_G|Y_G}}{P_{X_G}}}_{\substack{\text{Linear-quadratic} \\ \text{form in } X,Y}} = I(X_G;Y_G)$$

$$\Rightarrow \mathbb{E}_{X,Y}[\cdot] = \mathbb{E}_{X_G,Y_G}[\cdot]$$

$\Rightarrow$ WLOG optimize over jointly Gaussian

$Y \sim \mathcal{N}(0, \sigma^2) \perp\!\!\!\perp Z \sim \mathcal{N}(0,1)$

$X = \alpha Y + \beta Z$

$$I(X;Y) = \frac{1}{2}\log\left(1 + \frac{\alpha^2\sigma^2}{\beta^2}\right)$$

$$\mathbb{E}(X-Y)^2 = (1-\alpha)^2\sigma^2 + \beta^2 \geq \beta^2$$

$\Rightarrow$ setting $\sigma^2 \leftarrow \alpha^2\sigma^2$, $\alpha \leftarrow 1$

minimizes the bd

$$\Rightarrow \quad X = Y + \beta Z. \qquad \mathbb{E}(X-Y)^2 = \beta^2$$

$$\mathbb{E}Y^2 = 1 - \beta^2.$$

$$I(X;Y) = \frac{1}{2}\log\left(1 + \frac{1-\beta^2}{\beta^2}\right)$$

$$= \frac{1}{2}\log 1/\beta^2.$$

$$D_{vec}(R) = 2^{-2R}.$$

Opt-l $Y \sim \mathcal{N}(0, \underbrace{1-2^{-2R}}_{0.07})$

for VQ-GAN!

Projects: 1) training dynamics under STE (need to model Enc and Dec somehow. start with linear?)

2) Setup 3-MLP E and D and compare VQVAE, FSQ, Gumbel-Softmax for iid Gsn data.

+ Bottleneck dim size (esp. FSQ)

! 3) FSQ with commitment loss
+ $\|z - \hat{z}\|_2^2$ term?

FSQ + some VAE-style
regularization to enforce
uniform dist at bottleneck

(VAE has $z \sim \mathcal{N}(0, I_d)$
so just need to apply
$\Phi^{-1}(z)$ to get unif.)