

Inference with Low-Dimensional Structures

Druv Pai

UC Berkeley



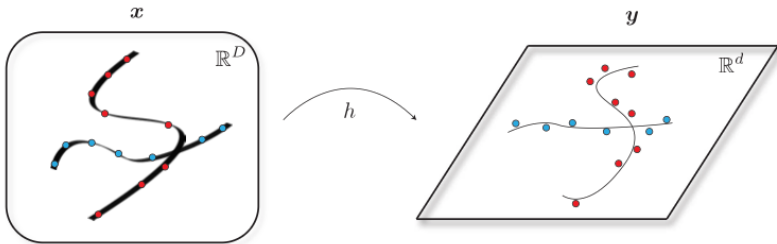
Lectures so Far

- History of the pursuit and study of intelligence.
- Learning and sampling via analytic methods, denoising, and/or compression.
- Objectives for representation learning: information gain.
- Novel white-box deep neural networks via optimization.

This lecture: inference!

Principled solutions for “downstream” tasks (prediction, completion, generation, etc.) using deep networks.

What is Inference?



Model problem:

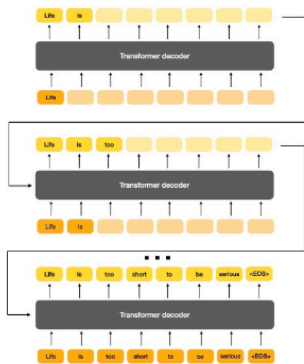
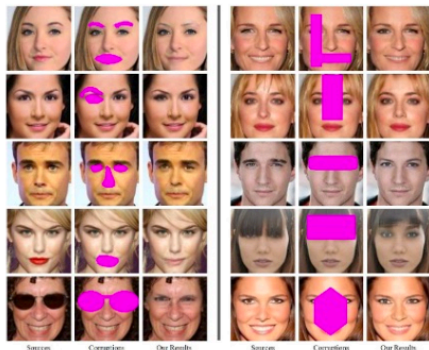
$$y = h(x) + w$$

- x : statistical/geometric low-dim. structure
- w : noise, sparse errors, missing values, etc.

Three main tasks:

- Find $\mathbb{E}[x \mid y = \nu]$
- Find $\arg \max_{\psi} p_{x|y}(\psi \mid \nu)$
- Sample from $p_{x|y}(\cdot \mid \nu)$

Examples of Inference



Examples: Regression, classification, denoising, completion (incl. next-token prediction), error correction, sampling, etc.

Inference From a Bayesian Perspective

The **posterior** $p_{x|y}$ is centrally important!

$$p_{x|y}(\psi \mid \nu) = \frac{p_{y|x}(\nu \mid \psi)p_x(\psi)}{p_y(\nu)} = \frac{p_{y|x}(\nu \mid \psi)p_x(\psi)}{\int_{\psi'} p_{y|x}(\nu \mid \psi')p_x(\psi') d\psi'}$$

- $p_{y|x}$: how well do we know (h, w) ?
- p_x : how well do we know x ?

MAP Estimation

Maximum a posteriori (MAP) estimation:

$$\begin{aligned} & \arg \max_{\psi} p_{\mathbf{x}|\mathbf{y}}(\psi \mid \boldsymbol{\nu}) \\ &= \arg \max_{\psi} \log p_{\mathbf{x}|\mathbf{y}}(\psi \mid \boldsymbol{\nu}) \\ &= \arg \max_{\psi} \{ \log p_{\mathbf{y}|\mathbf{x}}(\boldsymbol{\nu} \mid \psi) + \log p_{\mathbf{x}}(\psi) \} \end{aligned}$$

Compute MAP estimate via (e.g.) *gradient ascent*:

$$\psi \leftarrow \psi + \kappa [\nabla_{\psi} \log p_{\mathbf{y}|\mathbf{x}}(\boldsymbol{\nu} \mid \psi) + \nabla_{\psi} \log p_{\mathbf{x}}(\psi)].$$

Geometric Interpretations

Assume disturbance w is small (w.h.p.).

- Suppose x has (approx.) low-dimensional support \mathcal{S} .
- Let $F(\psi) := \min\{\|\psi - \psi'\|_2 : \psi' \in \mathcal{S}\}$ be “distance to \mathcal{S} ”.

A natural approach: Given $y = \nu$, try to recover x as:

$$\max_{\psi} \left\{ -\frac{1}{2} \|h(\psi) - \nu\|_2^2 \right\} \quad \text{s.t.} \quad F(\psi) = 0$$

Interpreting Augmented Lagrange Multipliers

Augmented Lagrange multiplier method:

$$\begin{aligned} & \max_{\boldsymbol{\psi}} \left\{ -\frac{1}{2} \|h(\boldsymbol{\psi}) - \boldsymbol{\nu}\|_2^2 + \boldsymbol{\lambda}^\top F(\boldsymbol{\psi}) - \frac{\mu}{2} \|F(\boldsymbol{\psi})\|_2^2 \right\} \\ &= \max_{\boldsymbol{\psi}} \left\{ -\frac{1}{2} \|h(\boldsymbol{\psi}) - \boldsymbol{\nu}\|_2^2 - \frac{\mu}{2} \left\| F(\boldsymbol{\psi}) - \frac{\boldsymbol{\lambda}}{\mu} \right\|_2^2 \right\} \end{aligned}$$

Two interpretations:

- **MAP estimation** with:

$$\begin{aligned} p_{\boldsymbol{x}}(\boldsymbol{\psi}) &\propto \exp \left\{ -\frac{\mu}{2} \left\| F(\boldsymbol{\psi}) - \frac{\boldsymbol{\lambda}}{\mu} \right\|_2^2 \right\} \\ p_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{\nu} \mid \boldsymbol{\psi}) &\propto \exp \left\{ -\frac{1}{2} \|h(\boldsymbol{\psi}) - \boldsymbol{\nu}\|_2^2 \right\} \end{aligned}$$

- **Constrained energy minimization** of quadratic forms.

Three Settings for Inference

- We know the distribution p_x ; we know the observation model $p_{y|x}$.
 - E.g., low-rank matrix completion.
- We have samples x from the distribution p_x ; we know the observation model $p_{y|x}$.
 - E.g., MAE, BERT, GPT.
- We have paired samples (x, y) from the distribution $p_{x,y} = p_x \cdot p_{y|x}$.
 - E.g., classification, text-conditioned image generation.

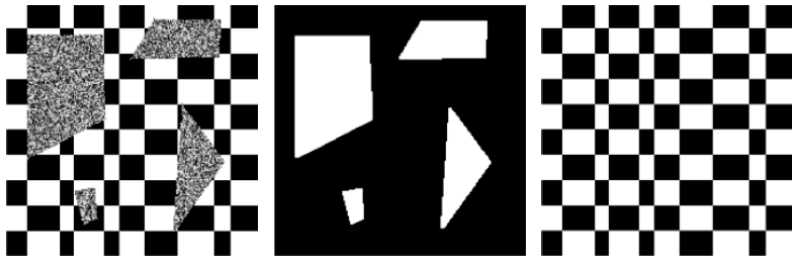
Vignette 1: Known Prior, Known Observation Model

Problem: *completion*.

- Observation model: $y = \mathcal{P}_\Omega(x)$
 - $\Omega \subseteq [n]$ is a set of indices.
 - $\mathcal{P}_\Omega(x)$ sets to 0 all entries of x indexed outside Ω .
- Goal: Given y , recover x .

Classical example: *low-rank matrix completion*.

- Prior: $x \in \mathbb{R}^{m \times n}$ has rank $r \ll \min(m, n)$.



Solving Low-Rank Matrix Completion

A natural idea: Given \mathbf{y} , find the *lowest rank* matrix agreeing with \mathbf{y} on the observed entries.

$$\min_{\mathbf{x}} \text{rank}(\mathbf{x}) \quad \text{s.t.} \quad \mathcal{P}_{\Omega}(\mathbf{x}) = \mathbf{y}.$$

Problem: Rank is *non-convex* and optimizing it is NP-hard.

Solution: Use a *convex relaxation* of the rank, the **nuclear norm** (sum of singular values). This is a *convex envelope* of the rank.¹

$$\min_{\mathbf{x}} \|\mathbf{x}\|_* \quad \text{s.t.} \quad \mathcal{P}_{\Omega}(\mathbf{x}) = \mathbf{y}.$$

Solve the associated Lagrangian minimization:

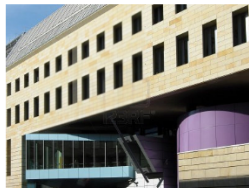
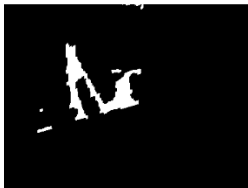
$$\min_{\mathbf{x}} \left\{ \|\mathbf{x}\|_* + \lambda \|\mathcal{P}_{\Omega}(\mathbf{x}) - \mathbf{y}\|_2^2 \right\}$$

¹Within the unit ball of the operator-norm unit ball.

Extensions to Low-Rank Matrix Completion

Low-rank plus sparse model for images:

$$y = \underbrace{h(x)}_{\text{distortion}} + \underbrace{w}_{\substack{\text{sparse} \\ \text{corruption/occlusion}}}$$



[Lia+12]

Known prior + known observation model
 \Rightarrow **bespoke efficient (non-deep) algorithms!**

Vignette 2: Samples from Prior, Known Observation Model

Same setup, except we only have finite samples $\{x_1, \dots, x_n\}$.

Question: How to obtain $\mathbb{E}[x \mid y = \nu]$, compute $\arg \max_{\psi} p_{x|y}(\psi \mid \nu)$, or sample from $p_{x|y}$?

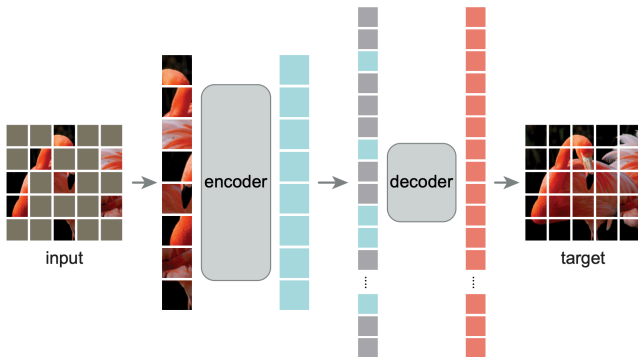
Train a (deep) model to complete x !

Natural examples:

- For image completion: **Masked Auto-Encoder (MAE)**
- For text completion: **Bidirectional Encoder Representations from Transformers (BERT)**
- In the special case where the last text token is corrupted: **Generative Pre-Trained Transformers (GPT)**

Masked Auto-Encoder

Train transformer encoder f and “decoder” g to jointly reconstruct x from $\mathcal{P}_{\Omega}(x)$.

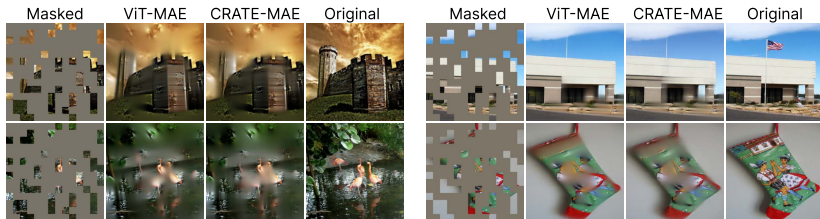


Loss:

$$\mathcal{L}_{\text{MAE}}(f, g) = \mathbb{E}_{\mathbf{x}, \Omega} \|\mathbf{x} - (g \circ f)(\mathcal{P}_{\Omega}(\mathbf{x}))\|_2^2$$

Results of MAE

Can it compute $\mathbb{E}[x \mid y]$? Yes!



[Pai+23]

$\mathbb{E}[x \mid y]$ = “average of all possible x resulting in corrupted y ”
= *not necessarily* a natural image!

How to sample from $p_{x|y}$ or compute $\arg \max_{\psi} p_{x|y}(\psi \mid \nu)$?

Diffusion Inspired Methodology

Diffusion-like methodology: *learn to denoise* as best as possible.

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \quad \mathbf{x}_t = \mathbf{x} + t\mathbf{g}, \quad t \in [0, T]$$

with $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and \mathbf{A} measurement operator.

Theorem ([Dar+23] Theorem 1): A denoiser $\bar{\mathbf{x}}^*$ where

$$\bar{\mathbf{x}}^*(t, \cdot, \cdot) \in \arg \min_{\bar{\mathbf{x}}(t, \cdot, \cdot)} \mathbb{E}_{\mathbf{x}, \mathbf{x}_t, \mathbf{y}} \|\mathbf{A}\bar{\mathbf{x}}(t, \mathbf{A}\mathbf{x}_t, \mathbf{A}) - \mathbf{y}\|_2^2$$

must satisfy

$$\mathbf{A}\bar{\mathbf{x}}^*(t, \mathbf{A}\mathbf{x}_t, \mathbf{A}) = \mathbf{A} \mathbb{E}[\mathbf{x} \mid \mathbf{A}\mathbf{x}_t, \mathbf{A}]$$

Note: If \mathbf{A} full (column) rank, then $\mathbb{E}[\mathbf{x} \mid \mathbf{A}\mathbf{x}_t, \mathbf{A}] = \mathbb{E}[\mathbf{x} \mid \mathbf{x}_t]$.

Upshot: Can (approx.) denoise up to measurement consistency.

Denoising With Partial Observations

Regular denoising process w/ learned denoiser:

$$\hat{\mathbf{x}}_{t_\ell} = \delta_\ell \hat{\mathbf{x}}_{t_{\ell-1}} + (1 - \delta_\ell) \bar{\mathbf{x}}^*(t_{\ell-1}, \mathbf{A} \hat{\mathbf{x}}_{t_{\ell-1}}, \mathbf{A})$$



Observations y

[Dar+23]

Samples $x \sim p_{x|y}$

Ambient diffusion ([Dar+23]) can sample *natural images* which are consistent with partial measurements!

Nonlinear Measurements?

Back to the model $\mathbf{y} = h(\mathbf{x}) + \mathbf{w}$. Want to find $\mathbb{E}[\mathbf{x} \mid \mathbf{y} = \boldsymbol{\nu}]$ or $\arg \max_{\boldsymbol{\psi}} p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{\psi} \mid \boldsymbol{\nu})$ or sample from $p_{\mathbf{x}|\mathbf{y}}(\cdot \mid \boldsymbol{\nu})$.

- Why not train separate denoisers for each instance of \mathbf{y} by using many noisy samples \mathbf{x}_t^ν corresponding to $\mathbf{y} = \boldsymbol{\nu}$?
- **Answer:** Prohibitively sample-inefficient, and unnecessary!

$$\begin{aligned} p_{\mathbf{x}_t^\nu}(\cdot) &= \int p_{\mathbf{x}_t^\nu | \mathbf{x}^\nu}(\cdot \mid \boldsymbol{\psi}) \cdot p_{\mathbf{x}^\nu}(\boldsymbol{\psi}) \, \mathrm{d}\boldsymbol{\psi} \\ &= \int p_{\mathbf{x}_t | \mathbf{x}, \mathbf{y}}(\cdot \mid \boldsymbol{\psi}, \boldsymbol{\nu}) \cdot p_{\mathbf{x} | \mathbf{y}}(\boldsymbol{\psi} \mid \boldsymbol{\nu}) \, \mathrm{d}\boldsymbol{\psi} \\ &= \int p_{\mathbf{x}_t, \mathbf{x} | \mathbf{y}}(\cdot, \boldsymbol{\psi} \mid \boldsymbol{\nu}) \, \mathrm{d}\boldsymbol{\psi} \\ &= p_{\mathbf{x}_t | \mathbf{y}}(\cdot \mid \boldsymbol{\nu}) \end{aligned}$$

\implies by Tweedie, $\mathbb{E}[\mathbf{x}^\nu \mid \mathbf{x}_t^\nu = \boldsymbol{\xi}] = \mathbb{E}[\mathbf{x} \mid \mathbf{x}_t = \boldsymbol{\xi}, \mathbf{y} = \boldsymbol{\nu}]$.

Learning with a Nonlinear Measurement Operator

Now how to compute $\mathbb{E}[x \mid x_t = \xi, y = \nu]$? Tweedie:

$$\mathbb{E}[x \mid x_t = \xi, y = \nu] = \xi + t^2 \nabla_{\xi} \log p_{x_t|y}(\xi \mid \nu)$$

Bayes:

$$\nabla_{\xi} \log p_{x_t|y}(\xi \mid \nu) = \underbrace{\nabla_{\xi} \log p_{x_t}(\xi)}_{\text{score}} + \underbrace{\nabla_{\xi} \log p_{y|x_t}(\nu \mid \xi)}_{\text{measurement}}$$

Combining them:

$$\mathbb{E}[x \mid x_t = \xi, y = \nu] = \mathbb{E}[x \mid x_t = \xi] + t^2 \nabla_{\xi} \log p_{y|x_t}(\nu \mid \xi)$$

Highly interpretable: Unconditional denoiser + measurement consistency correction!

Now: How to find the term $\nabla \log p_{y|x_t}$?

Finding the Posterior via DPS

Since \mathbf{y} and \mathbf{x}_t are conditionally independent given \mathbf{x} :

$$p_{\mathbf{y}|\mathbf{x}_t}(\boldsymbol{\nu} \mid \boldsymbol{\xi}) = \int p_{\mathbf{y}|\mathbf{x}}(\boldsymbol{\nu} \mid \boldsymbol{\psi}) p_{\mathbf{x}|\mathbf{x}_t}(\boldsymbol{\psi} \mid \boldsymbol{\xi}) \mathrm{d}\boldsymbol{\psi}$$

When t is small, $p_{\mathbf{x}|\mathbf{x}_t} \approx$ a Dirac delta around $\mathbb{E}[\mathbf{x} \mid \mathbf{x}_t]$. Hence

$$p_{\mathbf{y}|\mathbf{x}_t}(\boldsymbol{\nu} \mid \boldsymbol{\xi}) \approx p_{\mathbf{y}|\mathbf{x}}(\boldsymbol{\nu} \mid \mathbb{E}[\mathbf{x} \mid \mathbf{x}_t = \boldsymbol{\xi}]),$$

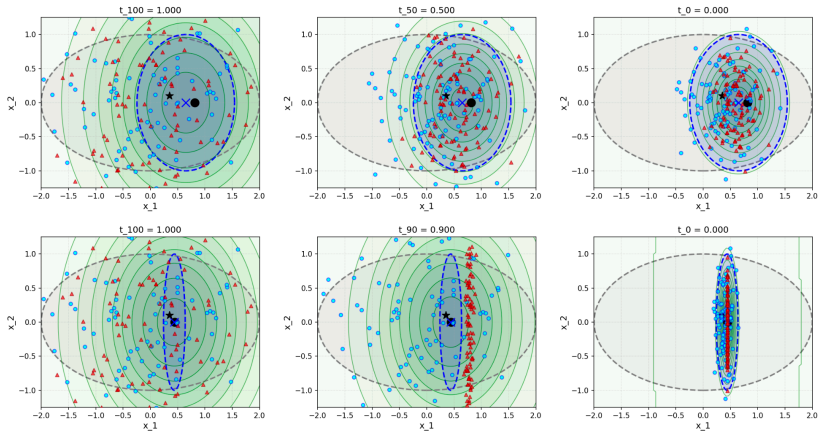
$$\nabla_{\boldsymbol{\xi}} \log p_{\mathbf{y}|\mathbf{x}_t}(\boldsymbol{\nu} \mid \boldsymbol{\xi}) \approx \nabla_{\boldsymbol{\xi}} \log p_{\mathbf{y}|\mathbf{x}}(\boldsymbol{\nu} \mid \mathbb{E}[\mathbf{x} \mid \mathbf{x}_t = \boldsymbol{\xi}]),$$

$$\mathbb{E}[\mathbf{x} \mid \mathbf{x}_t = \boldsymbol{\xi}, \mathbf{y} = \boldsymbol{\nu}] \approx \mathbb{E}[\mathbf{x} \mid \mathbf{x}_t = \boldsymbol{\xi}] + t^2 \nabla \log_{\boldsymbol{\xi}} p_{\mathbf{y}|\mathbf{x}}(\boldsymbol{\nu} \mid \mathbb{E}[\mathbf{x} \mid \mathbf{x}_t = \boldsymbol{\xi}]).$$

Note: by assumption we know $p_{\mathbf{y}|\mathbf{x}}(\boldsymbol{\nu} \mid \boldsymbol{\psi})$.

This is the **DPS approximation** ([Chu+22]).

Validity of DPS Approximation



DPS only works with *large* observation noise!

Summary of Vignette 2

Given samples of x and a known model $y = h(x) + w$, we can:

- Learn the unconditional denoiser $\mathbb{E}[x \mid x_t]$;
- Estimate the conditional denoiser $\mathbb{E}[x \mid x_t, y]$ w/ DPS;

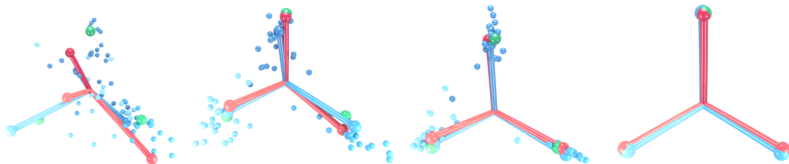
which enables:

- Conditional estimation $\mathbb{E}[x \mid y = \nu]$ (also w/ MAE, etc.);
- Conditional sampling from $p_{x|y}(\cdot \mid \nu)$;
- MAP estimation $\arg \max_{\psi} p_{x|y}(\psi \mid \nu)$.

Vignette 3: Paired Samples of Data and Observations

Model: $y = h(x) + w$. Both the prior p_x and observation model $p_{y|x}$ are *unknown*; instead have paired samples $\{(x_i, y_i)\}_{i=1}^N$.

First example: *classification*. y = label, x = data.



[PHD20]

y is a very compressed encoding of x !

Classification

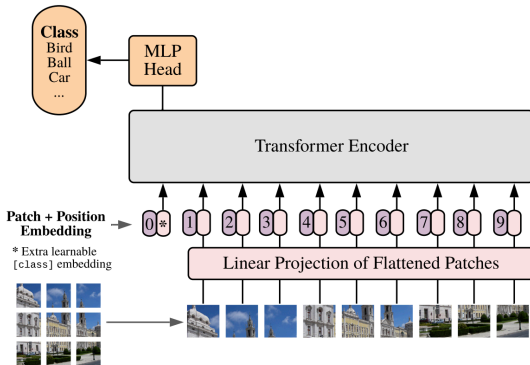
- $x \mapsto z$ (features) and $x \mapsto y$ (labels) are both compressive mappings, the latter more than the former.
- z captures the low-dim. structures in x , which y depends on. Hence $p_{z,y}$ is very structured/low-dim.
- Thus, plausible to learn a composite mapping $x \mapsto z \mapsto y$.
- Such mappings are built from compressive/denoising operators \implies learn a mapping $(x, x_{\text{cls}}) \mapsto (z, z_{\text{cls}}) \mapsto y$.

Motivates the **class token** in transformer classification.

Usually, deep classifiers use z to predict $p_{y|x}$ instead of just y .

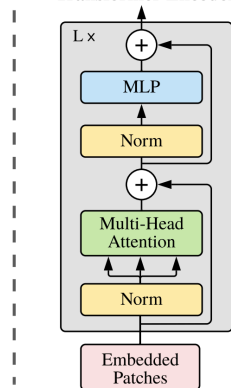
Classification in Transformers

Vision Transformer (ViT)



[Dos+20]

Transformer Encoder



Example 2: Class-Conditioned Image Generation

Setting: we have image-class pairs $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ and want to sample from $p_{\mathbf{x}|\mathbf{y}}(\cdot | \boldsymbol{\nu})$.

Naive Idea: Recall the starting point for conditional sampling:

$$\mathbb{E}[\mathbf{x} | \mathbf{x}_t = \boldsymbol{\xi}, \mathbf{y} = \boldsymbol{\nu}] = \mathbb{E}[\mathbf{x} | \mathbf{x}_t = \boldsymbol{\xi}] + t^2 \nabla_{\boldsymbol{\xi}} \log p_{\mathbf{y}|\mathbf{x}_t}(\boldsymbol{\nu} | \boldsymbol{\xi}).$$

Train a *denoiser* $\bar{\mathbf{x}}_{\theta_d}$ to estimate $\mathbb{E}[\mathbf{x} | \mathbf{x}_t]$ and a classifier $h_{\theta_c}: (t, \mathbf{x}_t) \mapsto [0, 1]^K$, problem solved?

No! Two issues:

- Classifier is too uncertain for moderate/large t
- Wasteful to train and use two models?

Classifier Guidance

Initial issue: Classifier h_{θ_c} is often very uncertain and its gradient is often too small to give good guidance.

Initial fix: Multiply the gradient by $\gamma \geq 1$.

Classifier Guidance Denoiser ([DN21]):

$$\bar{x}_{\theta}^{\text{CG}}(t, \xi, \nu) = \bar{x}_{\theta_d}(t, \xi) + \gamma t^2 \nabla_{\xi} \langle \log h_{\theta_c}(t, \xi), \nu \rangle$$

If all denoisers are perfectly learned:

$$\bar{x}^{\text{CG,ideal}}(t, \xi, \nu) = (1 - \gamma) \mathbb{E}[x \mid x_t = \xi] + \gamma \mathbb{E}[x \mid x_t = \xi, y = \nu].$$

No longer (approximating) a conditional expectation!

Classifier-Free Guidance

Secondary issue: Too wasteful, and empirically worse, to train two denoisers.

Solution: Use *the same* denoiser network \bar{x}_θ to estimate both conditional expectations.

- Introduce a new pseudo-label \emptyset which means “no class”.
- Let $M(\nu)$ mask ν to \emptyset with some prob. (≈ 0.2), and train:

$$\mathcal{L}_{\text{CFG}}(\theta) := \mathbb{E}_t w(t) \mathbb{E}_{\mathbf{x}, \mathbf{x}_t, \mathbf{y}, M} \|\mathbf{x} - \bar{\mathbf{x}}_\theta(t, \mathbf{x}_t, M(\mathbf{y}))\|_2^2$$

Classifier-Free Guidance Denoiser ([HS22])

$$\bar{\mathbf{x}}_\theta^{\text{CFG}}(t, \boldsymbol{\xi}, \boldsymbol{\nu}) = (1 - \gamma) \bar{\mathbf{x}}_\theta(t, \boldsymbol{\xi}, \emptyset) + \gamma \bar{\mathbf{x}}_\theta(t, \boldsymbol{\xi}, \boldsymbol{\nu}).$$

Classifier-Free Guidance Examples



$\gamma = 0.0$

$\gamma = 3.0$

Text-Conditioned Generation

What if instead of *labels*, we had *text prompts* y ?

Goal: Still to sample from $p_{x|y}$.

Solution: Encode y into features, and train a denoiser for CFG.



Summary

Discussed inference in three representative scenarios:

- Know both the prior and observation model;
- Have samples from the prior (data) and know the observation model;
- Have paired samples of data and observations.

Lots of relevant topics in our book that couldn't fit:

- Cross-attention and architectural choices
- Principles for inference when we *only* have observations;
- Principles for distributed inference, etc.

References I

- [Chu+22] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, et al. “Diffusion posterior sampling for general noisy inverse problems”. In: *arXiv preprint arXiv:2209.14687* (2022).
- [Dar+23] Giannis Daras, Kulin Shah, Yuval Dagan, et al. “Ambient diffusion: Learning clean distributions from corrupted data”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 288–313.
- [DN21] Prafulla Dhariwal and Alexander Nichol. “Diffusion models beat gans on image synthesis”. In: *Advances in neural information processing systems* 34 (2021), pp. 8780–8794.

References II

- [Dos+20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).
- [He+22] Kaiming He, Xinlei Chen, Saining Xie, et al. “Masked autoencoders are scalable vision learners”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 16000–16009.
- [HS22] Jonathan Ho and Tim Salimans. “Classifier-free diffusion guidance”. In: *arXiv preprint arXiv:2207.12598* (2022).

References III

- [Lia+12] Xiao Liang, Xiang Ren, Zhengdong Zhang, et al. “Repairing sparse low-rank texture”. In: *European Conference on Computer Vision*. Springer. 2012, pp. 482–495.
- [Pai+23] Druv Pai, Ziyang Wu Wu, Sam Buchanan, et al. “Masked completion via structured diffusion with white-box transformers”. In: *International Conference on Learning Representations*. 2023.
- [PHD20] Vardan Papyan, XY Han, and David L Donoho. “Prevalence of neural collapse during the terminal phase of deep learning training”. In: *Proceedings of the National Academy of Sciences* 117.40 (2020), pp. 24652–24663.