# Questions

## Written

### Classical

- General Statistics
  - What is the definition of the Moore-Penrose pseudoinverse and how does one calculate it?
  - What is the $L^p, L^\infty$ and Frobenius norm?
  - What is the connection between eigenvalues and determinant/trace?
  - What is the Jacobian, and what is a Hessian?
  - Write down taylor's theorem, and the 2nd order multivariate approximation of a function.
  - What is the relationship between Hessian, eigenvalues and stationary points?
  - Write down the multivariate change of variables formula. How does this relate to integration / pdfs?
  - Write down the primal and dual form for the generalized lagrangian, the minimax objectives, and the KKT conditions
  - What is the formula for conditional expectation and the law of iterated expectation?
  - What's the formula for $\text{Cov}(f(X), g(Y))$?
  - What's the formula for $\mathbf{\Sigma} \stackrel{\text{def}}{=} \text{Cov}_{\mathbf{x}\sim p}[\mathbf{x}]$? Formula for correlation? Why is this psd?
  - Suppose $X_i$ are iid and $E[X] = \mu$ and $\text{Var}(X) = \sigma^2$.
    - What's an unbiased estimate of $\sigma^2$?
    - What's the variance of this estimate? (Hint: What's its distribution?)
  - Write down two formulae for the sample covariance matrix $\mathbf{S}$. What is $S_{ij}$?
  - What's the CLT? LLN?
  - What's the MGF/CF? Calculate the MGF of a univariate normal.
  - What is Self-Information of an event, Entropy, Information Gain, Cross-Entropy, KL Divergence, Gini Impurity, Perplexity?
  - What is generalization error?
  - Decompose MSE in both the model prediction case and the parameter estimation case.
  - What are 4 reasons to use parameter regularization?
  - For a quadratic loss function, how do parameter estimates change for L2 regularization? What's the geometric intuition for this?
  - For a quadratic loss function with diagonal Hessian, how do parameter estimates change for L2 regularization? What's the geometric intuition for this?
  - What is a consistent estimator? Give an example of a consistent but biased estimator. What about an unbiased by inconsistent estimator?
  - What is the relationship between MLE, KL-divergence, CE, and MSE?
  - What are the two formulae for Fisher's Information? What does it mean for an estimator to be efficient?
  - What is the relationship between MLE, consistency and efficiency? Why is regularization useful?
  - Write out formulae for Sensitivity, Specificity, Accuracy, Precision, Recall, FPR, F1 score.
  - Write down formulae for CE, NCE, Huber loss, quantile loss, hinge loss, MAPE, Symmetric MAPE, MRR, mAP, nDCG, FID, Inception Score, FAD, BLEU, InfoNCE.
  - What is Alpha and Beta in a test? Significance level, confidence level, power?
  - What is a Z, T, Chi-squared test, ANOVA? What are the different types of t-tests?
  - What is Multiple Hypothesis Testing and what are ways to correct for it?
  - What are the formulae for AIC and BIC?
  - What is the probability distribution of a Poisson, Geometric, Exponential, Normal and Multivariate Normal distribution?
  - What is the mean and variance of a Binomial, Poisson, Geometric, Exponential, Beta, Gamma, Chi-Squared distribution?
  - What is the formula for a Jeffrey's Prior? What does invariance mean? What's the intuition for it?
  - Describe Gibbs, MH, HMC, LMC, Rejection, Importance sampling.
  - What is the taylor expansion for $e^x$?.
- Linear Regression
  - What are the model specifications and associated assumptions?
  - What are tests for assumptions and remedies?
  - What is the MLE estimate for $\mathbf{B}$? In 2D? With weights?

  - What is the MLE and unbiased estimate for $\sigma^2$?
  - What is the gradient and hessian of the loss function with respect to $\mathbf{B}$?
  - If $\sigma^2$ is known, what is the distribution of $\hat{\mathbf{B}}$? How does SSE relate to $\sigma^2$?
  - What is the MSE of $\hat{\mathbf{B}}$?
  - What is the formula of a t-stat and why is it low with multicollinearity?
  - What is the formula for $R^2$, adjusted $R^2$ and that of the sample Pearson correlation coefficient?
  - What is the chow test?
- Naive Bayes, Logistic Regression and GLMs
  - What is the model for Naive Bayes and what are its assumptions?
  - What are the model specifications for Logistic Regression and what are its assumptions?
  - What is the loss for Logistic Regression? What about softmax regression?
  - What is the gradient and hessian of the loss function with respect to $\mathbf{B}$? What about softmax regression?
  - Compare Naive Bayes to Logistic Regression.
  - What is the formula for GLMs?
- SVMs
  - What are the model specifications and associated assumptions?
  - Give a form that the hyperplane takes.
  - Write down the equation that needs to be optimized to solve the problem. What about soft-margin SVMs?
  - What is the formula for kernel regression? How do you apply the kernel trick to SVMs?
  - How do you change the formula for support vector regressions?
  - How does changing $C$ change the results of an SVM in `sklern`?
- Decision Trees
  - Write down the loss used for node splitting.
  - Write down the loss used by CART for pruning.
  - Is a node's Gini impurity generally/always lower or higher than its parent's?
- Ensemble Learning
  - What is bagging/pasting and how is it used in RF?
  - What are two ways of calculating feature importance in an RF?
  - What are two boosting algorithms?
  - Write the general formulae of GAMs
  - Name features about XGBoost and LightGBM
  - Compare Boosting with RFs
- Dimensionality Reduction
  - How would you combat the curse of dimensionality?
  - Write down the PCA algorithm
  - Describe the algorithm for LLE and t-SNE
  - In what cases would you use regular PCA, incremental PCA, randomized PCA, or random projection?
  - PCA vs LDA
  - t-SNE vs UMAP?
- Unsupervised Clustering
  - Describe K-means, DBSCAN, Spectral Clustering, Agglomerative Clustering
  - Write down the model specification, Auxiliary Function and update steps of a GMM
  - What are metrics used in clustering?
  - Discuss two clustering algorithms that can scale to large datasets
- Gaussian Process
  - Write down the model specification of the GP prior, joint likelihood, conditional likelihood
- Causal Inference
  - What are key assumptions in Causal Inference?
  - Categorize the type of experiments we encounter
  - Write down both the Outcome Modeling and IPW estimates for ATE in an Observational Study.
  - When would you use a regression, 2-step least squares, or Double ML?
- ARIMA
  - Write down the model for an $AR, MA, ARMA, ARIMA, SARIMA$ model
  - What are the assumptions of the ARIMA model?

**DL**

- Basics
  - What's the difference between `binary_cross_entropy_with_logits` vs `cross_entropy`?
  - Explain the difference between `view` and `reshape`
  - Write down equations for Batch and Layer normalization. What is their function? Why would you use one over the other?
  - Why would you use PyTorch buffers?
  - What is PyTorch's default initialization for linear layers?
- Activations
  - What is the formula of sigmoid, tanh, arctan, relu, leaky relu, softplus, elu, gelu, silu?
  - What issues are faced by sigmoid/tanh and relu activations? Compare the two.
- Initialization
  - Explain why permutation invariance of networks can be an issue
  - Derive Xavier and Kaiming initialization
- Optimization
  - Name 7 ways of regularization
  - What are 3 issues with GD?
  - Under the quadratic approximation of loss, what's the optimal learning rate or and associated convergence rate?
  - Derive Newton's method
  - What are two issues with Newton's method? What can we do about these issues?
  - Explain the algorithm for momentum, adagrad, RMSProp, Adadelta, Adam. Compare SGD, SGDM and Adam for different loss function contours.
  - How does the convergence rate change for momentum?
  - How do you deal with the vanishing gradient problem?
- CNNs
  - How do shapes change after a convolutional kernel? How many parameters does this have? What's the time complexity of this?
  - Describe changes from LeNet, AlexNet, VGG, NiN, Inception, ResNet, DenseNet, U-Net. Touch on depth-wise separable layers.
  - ResNet architectures: pre vs post activation, ResNet-D, ResNeXt
- RNNs
  - Write down the equations for RNNs, Bidirectional RNNs, LSTMs (+ Peephole connections), GRUs
  - What are 3 issues for RNNs? How do we address them?
- Attention and Transformers
  - Draw the architecture for encoder-only, encoder-decoder, and decoder-only transformers
  - Write down the equations for attention
  - Write down the equations for different types of embeddings
  - What is the intuition for MLP layers?
  - Describe how Q and K composition work for the creation of induction heads
  - Name 3 ways that we have sped up implementation
  - Pre/Post-layer normalization, GeLU/SwiGLU, LayerNorm/RMSNorm
  - What are the advantages of transformers over traditional sequence-to-sequence models?
- VAEs
  - Derive the training loss for VAEs
- Diffusion
  - Derive the training loss for DDPM
  - Explain the intuition for different variance schedules
  - Draw the U-Net for Stable Diffusion
  - Explain classifier-free guidance
  - Draw the architecture for LDMs
  - Describe faster sampling techniques and derive the sampling step for DDIM
  - Without incorporating textual conditioning in training, how might we use guidance to enable textual conditioning in inference?
  - What are 2 ways to condition on both image and text without additional training? Now give 2 ways to do so with fine-tuning.

- Flows
  - Write down the likelihood for normalizing flow
  - Describe the architecture of RealNVP, GLOW and autoregressive flows.
  - Describe residual flows.
  - Describe continuous normalizing flows. Write down the continuity/transport equation and the log likelihood.
  - Describe flow matching. How is this related to diffusion?
  - What is dequantization and variational dequantization?
  - What is the multi-scale architecture?
- GANs
  - Describe 2 types of white box attacks. What can we do abotu these?
  - Name 4 issues with GANs?
  - What 5 tricks to improve GAN training?
  - Describe 3 types of GANs
- GNNs
  - What kind of problems can GNNs solve?
  - What's the intuition behind GNNs, and how does this relate to CNNs?
  - Describe the architectural differences between GCN, GraphConv and Graph Attention Learning
  - Describe the 3 approaches to meta learning with examples
  - Describe 3 models trained with contrastive learning
  - What is active learning?
- Computer Vision
  - Autoregressive modeling: Describe 2 models for generation and 1 for classification
  - What's the difference between segmentation and object detection?
  - How does object detection usually work?
  - What's the difference between R-CNN, Fast R-CNN, Faster R-CNN and Mask R-CNN?
  - Describe YOLO, SSD and RetinaNet
  - How would you create a 3D model of an object from imagery and depth sensor measurements taken at all angles around the object?
- NLP
  - What is tokenization, normalization, pre-tokeniation, stemming, lemmatization
  - Describe 4 subword tokenization algorithms
  - Explain tf-idf
  - What are 3 algorithms for generating context-independent token embeddings?
  - What are 2 models for generating context-dependent token embeddings?
  - Describe how BERT and T5 are trained
  - Describe architectural differences between various decoder-only models
  - What are 4 ways we can fine-tune BERT for various tasks?
  - What are 3 ways to guide pre-trained models, and 4 ways to fine-tune them?
  - What are components of an LLM agent?
- RL
  - Write down the Bellman equations for value and action-value function
  - What are the differences between Value/Q-based and policy-based learning?
  - For Q-learning, what are two training methodologies we can employ to make parameter updates?
  - What are on-policy and off-policy methods?
  - What is Deep Q-Learning? What are 3 tricks we use to stabilize training?
  - What is Proximal Policy Learning?
- Audio
  - What audio tokenizers are there? How are they usually trained and consequently different?
  - Name 5 diffusion-based models and 4 autoregressive models and describe their architecture and any relevant details
  - What research has been done regarding controllability of music generation? What are you excited about?
  - How have agents been applied in music generation? What are you excited about?
  - How has domain knowledge been used in music generation? What are you excited about?
- Multimodal
  - How does NExT-GPT work? Give examples of each component and training objectives.

- Post-Training
  - How might be reduce latency or storage at the cost of performance?
  - What is guidance and what does this look like for NLP?
  - What is fine-tuning and what are 3 methods of PEFT?
  - What is the process for RLHF? DPO? IPO?
- Hyperparameter optimization
  - What are the components of a HPO library?
  - What is an example of multi-fidelity hyperparameter optimization? How can we make this faster?
- Computational performance
  - What are differences between a CPU and a GPU
  - What is multithreading vs multiprocessing?
  - What is vectorization?
  - What is imperative vs symbolic programming?
  - What are features of JAX?
  - What at common bottlenecks and solutions to them?
  - What is DP and FSDP? What can be sharded in FSDP? What is the difference between `DP` and `DDP`? Elaborate on synchronous computation.
  - What is an issue with pipeline parallelism and how can we remedy this?
  - What is gather and scatter and when would we use these? How can we leverage asynchronous layers to reduce compute time?
  - What is disaggregated serving? When might we want to avoid this?

**SD + Coding Practices**

- System Design
  - Write down the steps/template for a typical SWE SD interview
    - Explain each component
  - Write down the steps/template for a typical ML SD interview
  - Run through a template recommender system SD interview. Review past architectures used for this problem.
  - What's the difference between collaborative filtering and content-based filtering?
  - Name some common online and offline metrics that can be used for each use case
  - SQL vs No-SQL.
  - What is an ACID database?
  - What is the CAP theorem?
- Modeling
  - How would you deal with outliers?
  - How can you deal with data imbalance?
  - Given a left-skewed distribution that has a median of 60, what conclusions can we draw about the mean and the mode of the data?
  - What are the different ways that data is missing?
  - How would you deal with missing data?
  - What are different types of distributional drift?
- OOP / Python
  - How is garbage collection done for Python/Java?
  - What are the principles of OOP?
  - What is the difference between an iterator and a generator?
  - What are decorators? Give examples of these.
  - What are `*args` and `*kwargs`?
  - How to initialize a 2d array in python? What about nested list comprehension?
  - Explain the rules for variable scope?
  - What are positional, keyword, and default arguments to a function?
  - Explain trailing and leading underscores in Python.
  - Explain `if __name__ == "__main__"`
  - Difference between lists, arrays and sets?

## Code

- OOP: Design an LRU Cache, Text Editor, (Ultimate) Tic Tac Toe, DenseNet
- Code up a basic flask app to describe RESTful APIs
- Code up various classical algorithms with `sklearn`, including code for data loading, model initialization, training and evaluation.
- Code up linear/logistic regression as a neural network, including code for data loading, model specification, initialization, and training.
- Code up various activation functions
- Derive and code up various initialization functions
- Derive and code up various optimization functions
- Code up a self/cross-attention head
- Code up GPT-2. With/without custom initialization, dropout, bias, RoPE, grouped attention, and LayerNorm/RMSNorm, pre/post-layer norm, gelu, swiglu
- Code up the architecture for an encoder-decoder model
- Code and analyze Kahn, Dijkstra, Bellman Ford, Floyd Warshall, Kruskal, Prim, Ford Fulkerson, Kosaraju, Manacher, Union Find