VARIATIONAL INFERENCE

BST 228

Speaker notes

No notes on this slide.

WELCOME BACK & TODAY

- Variational approaches for scalable inference.
- Theoretical background.
- Application to recommender systems and single-cell RNA sequencing data.
- Scaling to massive datasets with stochastic variational inference.

Speaker notes

No notes on this slide.

THE STANDARD PROBLEM

$p(\boldsymbol{\theta} \mid \mathbf{y})$ is not tractable for parameters $\boldsymbol{\theta}$ and data $\mathbf{y}$.

Speaker notes

No notes on this slide.

SOLUTIONS

- Manual Gibbs or Metropolis-Hastings sampler.
- Probabilistic programming languages with general-purpose samplers, such as JAGS or Stan.
- *Variational inference.*

Speaker notes

No notes on this slide.

VARIATIONAL INFERENCE

We approximate $p(\boldsymbol{\theta} \mid \mathbf{y}) \approx q(\boldsymbol{\theta})$ by a more convenient distribution q.

Speaker notes

- “Convenient” means anything that makes our lives easier, e.g., we could use a normal distribution to approximate the posterior if it has a single mode.
- More sophisticated but still tractable approximations include, for example, Gaussian mixture models and normalizing flows.

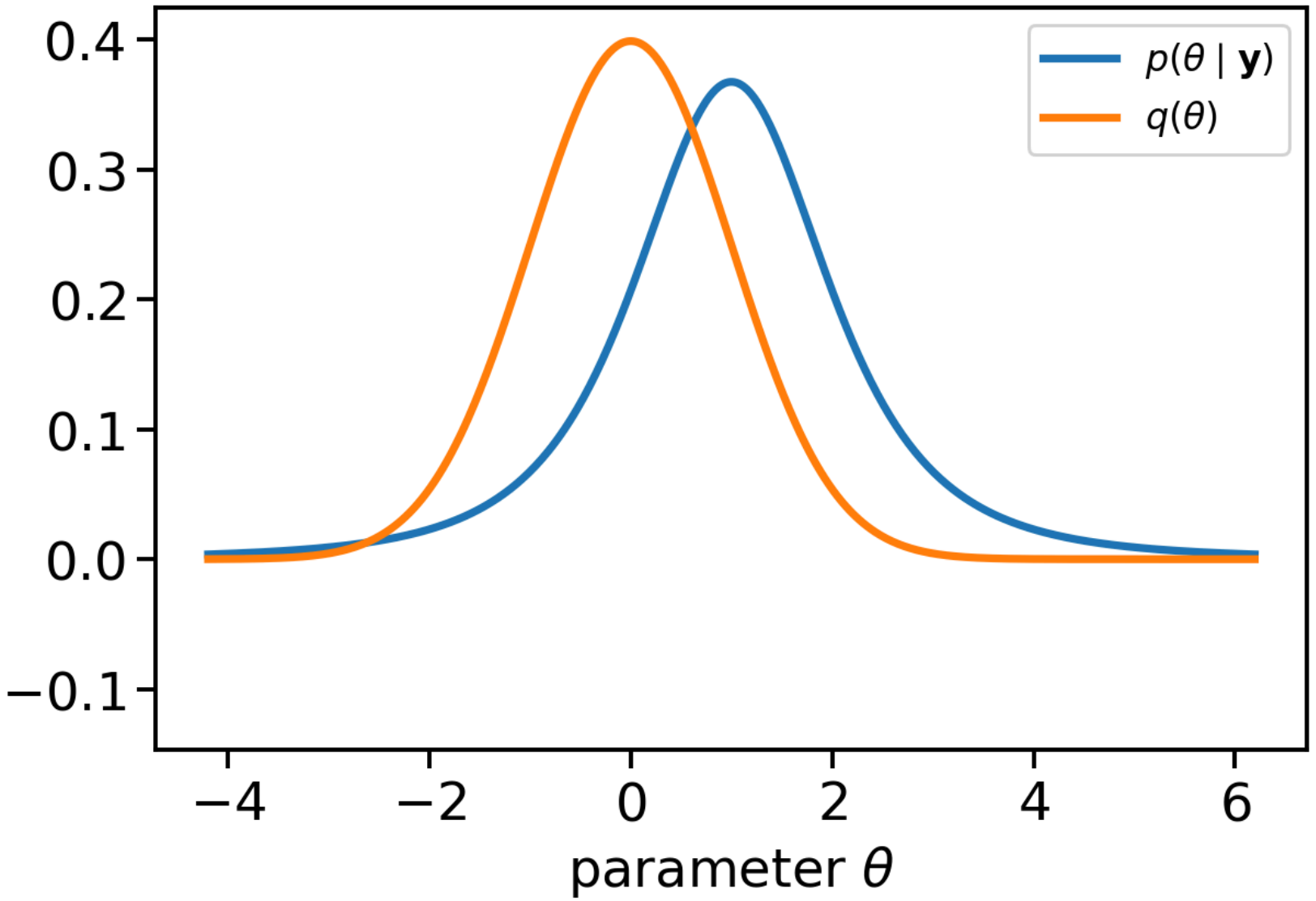
WHAT DOES $\approx$ MEAN?

We want to minimize the Kullback-Leibler divergence

$$D\left[q\left(\boldsymbol{\theta}\right)\|p\left(\boldsymbol{\theta}\mid\mathbf{y}\right)\right]=\int d\boldsymbol{\theta}\, q\left(\boldsymbol{\theta}\right)\log\left(\frac{q\left(\boldsymbol{\theta}\right)}{p\left(\boldsymbol{\theta}\mid\mathbf{y}\right)}\right).$$

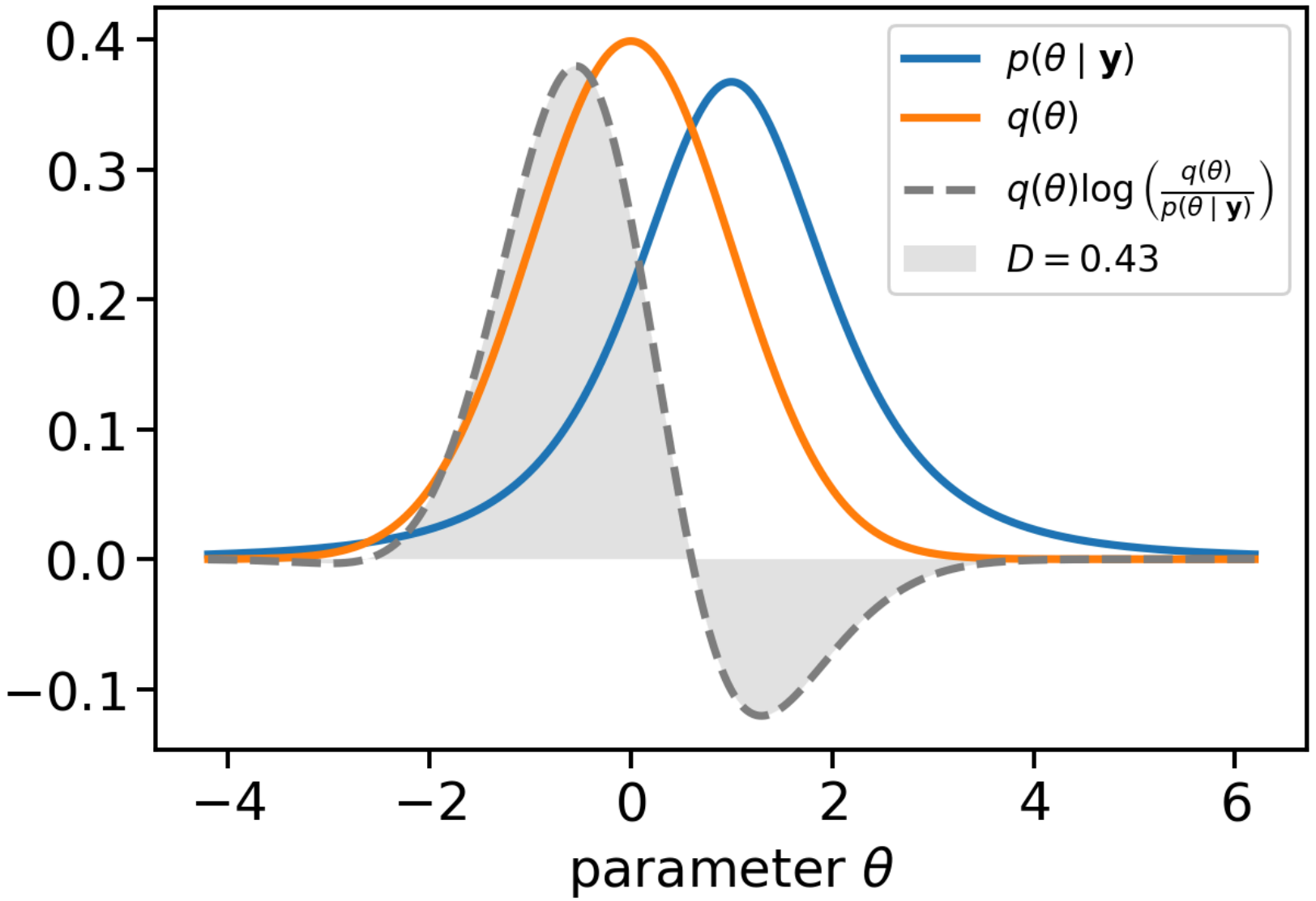
Speaker notes

- The KL divergence is a discrepancy measure between distributions. It is zero if and only if the two distributions are exactly the same.
- It is *not* a distance because it is not symmetric. Changing the order of the two distributions changes the result.
- Minimizing the KL divergence with respect to q prioritizes that q has small density where p has small density.
- Other discrepancy measures are reasonable, e.g., the reversed KL divergence, leading to expectation propagation.



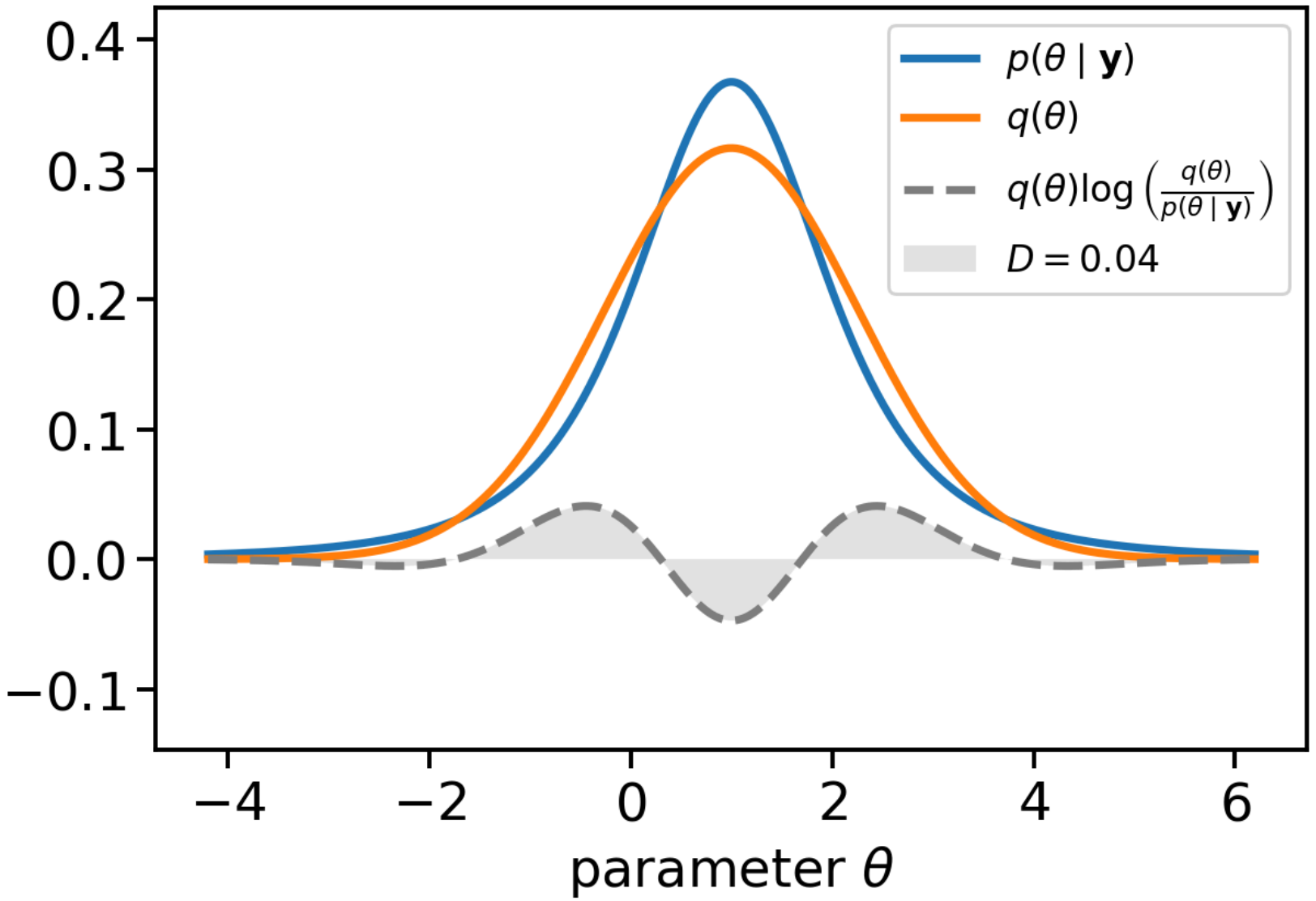
Speaker notes

- Shown in blue, $p\left(\theta \mid \mathbf{y}\right)$ is the true posterior for a univariate parameter θ and data $\mathbf{y}$. In this simple example, we assume a Student-t distribution with three degrees of freedom.
- Shown in orange, $q\left(\theta\right)$ is a Gaussian approximation whose parameters we will optimize to approximate the true posterior well. The distribution shown is the unoptimized standard normal distribution.



Speaker notes

- The dashed line is the integrand $q\left(\boldsymbol{\theta}\right)\log\left(\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})}\right)$ of the Kullback-Leibler divergence.
- The shaded area represents the Kullback-Leibler divergence.



Speaker notes

- Minimizing the Kullback-Leibler divergence results in a close approximation of the Gaussian variational approximation to the Student-t distribution which is our target.
- The Kullback-Leibler divergence for the optimized approximation is *much* smaller: 0.04 vs 0.43 for the unoptimized standard normal distribution.

SO WHY *VARIATIONAL*?

- The Kullback-Leibler divergence is a functional, i.e., a map from functions $\mathcal{Q}$ to $\mathbb{R}$.
- Calculus of *variations* considers minima of functionals and is the right tool to find an optimal approximation q^*.
- Optimization is a much easier task than sampling from an intractable distribution.

Speaker notes

No notes on this slide.

MINIMIZING D (1 / 3)

If we don't know $p(\boldsymbol{\theta} \mid \mathbf{y})$, how can we minimize

$$D\left[q(\boldsymbol{\theta}) \| p(\boldsymbol{\theta} \mid \mathbf{y})\right] = \int d\boldsymbol{\theta}\, q(\boldsymbol{\theta}) \log\left(\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \mathbf{y})}\right)?$$

Speaker notes

- To minimize the KL divergence (or an equivalent expression), we need to express it in terms of quantities we can evaluate.

MINIMIZING D (2 / 3)

$$
\begin{aligned}
q^* &= \underset{q \in \mathcal{Q}}{\arg\min}\, D\left[q\left(\boldsymbol{\theta}\right) \| p\left(\boldsymbol{\theta} \mid \mathbf{y}\right)\right] \\
&= \underset{q \in \mathcal{Q}}{\arg\min} \int d\boldsymbol{\theta}\, q\left(\boldsymbol{\theta}\right) \log\left(\frac{q\left(\boldsymbol{\theta}\right)}{p\left(\boldsymbol{\theta} \mid \mathbf{y}\right)}\right) \\
&= \underset{q \in \mathcal{Q}}{\arg\min} \int d\boldsymbol{\theta}\, q\left(\boldsymbol{\theta}\right) \log\left(\frac{q\left(\boldsymbol{\theta}\right) p\left(\mathbf{y}\right)}{p\left(\boldsymbol{\theta} \mid \mathbf{y}\right) p\left(\mathbf{y}\right)}\right)
\end{aligned}
$$

Speaker notes

- The second line follows from the definition of the Kullback-Leibler divergence.
- The third line follows from multiplying both nominator and denominator by the marginal likelihood $p\left(\mathbf{y}\right)$.

MINIMIZING D (3 / 3)

$$
\begin{aligned}
q^* &= \arg\min_{q \in \mathcal{Q}} \int d\boldsymbol{\theta}\, q(\boldsymbol{\theta}) \log \left(\frac{q(\boldsymbol{\theta})\, p(\mathbf{y})}{p(\boldsymbol{\theta} \mid \mathbf{y})\, p(\mathbf{y})} \right) \\
&= \arg\min_{q \in \mathcal{Q}} \int d\boldsymbol{\theta}\, q(\boldsymbol{\theta}) \left[\log q(\boldsymbol{\theta}) + \log p(\mathbf{y}) \right. \\
&\qquad\qquad \left. - \log p(\boldsymbol{\theta}, \mathbf{y}) \right] \\
&= \arg\min_{q \in \mathcal{Q}} \int d\boldsymbol{\theta}\, q(\boldsymbol{\theta}) \left[\log q(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}, \mathbf{y}) \right]
\end{aligned}
$$

Speaker notes

- The first line is copied from the previous slide.
- The second line follows from noting that the denominator is the joint distribution $p(\boldsymbol{\theta}, \mathbf{y})$ and expanding the $\log$.
- The third line follows from dropping the marginal likelihood because it does not depend on the approximation q.
- In principle, we can now evaluate all parts of the integrand.

VARIATIONAL LOSS

The loss functional to minimize is

$$L\left[q\left(\boldsymbol{\theta}\right)\right] = \int d\boldsymbol{\theta}\; q\left(\boldsymbol{\theta}\right)\left[\log q\left(\boldsymbol{\theta}\right) - \log p\left(\boldsymbol{\theta}, \mathbf{y}\right)\right].$$

But this still requires evaluating an intractable integral, and optimizing functionals to find optimal functions is difficult.

Speaker notes

No notes on this slide.

CALCULUS OF VARIATIONS TO "NORMAL" CALCULUS (1 / 3)

We use an approximation $q(\boldsymbol{\theta}; \boldsymbol{\phi})$ from a parametric family $\mathcal{Q}'$ with parameters $\boldsymbol{\phi}$.

In our example, $\mathcal{Q}'$ is the set of all normal distributions with parameters $\boldsymbol{\phi} = \{\mu, \sigma^2\}$.

Speaker notes

No notes on this slide.

CALCULUS OF VARIATIONS TO "NORMAL" CALCULUS (2 / 3)

Then

$$\arg\min_{q \in \mathcal{Q}'} L\left[q\left(\boldsymbol{\theta}; \boldsymbol{\phi}\right)\right] \Longleftrightarrow \arg\min_{\boldsymbol{\phi} \in \mathbb{R}^{\bullet}} L\left[q\left(\boldsymbol{\theta}; \boldsymbol{\phi}\right)\right],$$

where $\mathcal{Q}'$ is the chosen parametric family.

Speaker notes

- Optimizing over distributions in the family $\mathcal{Q}'$ is equivalent to optimizing the parameters of these distributions. The latter approach is both easier and more familiar: We "only" need to optimize a set of parameters $\boldsymbol{\phi}$.
- Note that the parameters $\boldsymbol{\phi}$ are *not* model parameters. They are parameters of the variational approximation and do not appear in the model definition.

CALCULUS OF VARIATIONS TO "NORMAL" CALCULUS (3 / 3)

We can evaluate gradients $\frac{\partial}{\partial \boldsymbol{\phi}} L\left[q\left(\boldsymbol{\theta}; \boldsymbol{\phi}\right)\right]$ and use our favorite optimization algorithm to find optimal parameters $\boldsymbol{\phi}^*$.

Speaker notes

No notes on this slide.

VARIATIONAL INFERENCE RECAP

- Approximate posterior $p\left(\boldsymbol{\theta} \mid \mathbf{y}\right)$ by a simpler distribution $q\left(\boldsymbol{\theta}\right)$.
- Minimizing the Kullback-Leibler divergence $D\left[q\left(\boldsymbol{\theta}\right) \| p\left(\boldsymbol{\theta} \mid \mathbf{y}\right)\right]$ is equivalent to minimizing the variational loss $L\left[q\left(\boldsymbol{\theta}\right)\right]$.
- For $q\left(\boldsymbol{\theta}\right)$ from a parametric family $\mathcal{Q}'$ with parameters $\boldsymbol{\phi}$ such that $L\left[q\left(\boldsymbol{\theta}; \boldsymbol{\phi}\right)\right]$ is tractable, we can find optimal parameters $\boldsymbol{\phi}^*$.
- Once optimized, we can draw an arbitrary number of independent samples from q^*.

Speaker notes

No notes on this slide.

RECOMMENDER SYSTEMS

- We have a matrix of ratings $\mathbf{Y} \in \mathbb{R}^{n \times m}$ for $n \sim 10^5$ users and $m \sim 10^4$ with 10^6 non-missing elements (see MovieLens 32M dataset).
- We want to predict missing ratings and recommend movies that users are likely to rate highly.
- This is the famous Netflix Prize with $1m prize money in 2009.
- The same model is often used for single-cell RNA sequencing data: users are cells and movies are genes.

Speaker notes

No notes on this slide.

LATENT FACTOR MODEL FOR RECOMMENDATIONS (1 / 2)

- Mean ratings μ.
- User satisfaction $\mathbf{a} \in \mathbb{R}^n$.
- User embeddings $\mathbf{A} \in \mathbb{R}^{n \times r}$ in r dimensions for each user.
- Movie quality $\mathbf{b} \in \mathbb{R}^m$.
- Movie embeddings $\mathbf{B} \in \mathbb{R}^{m \times r}$.
- Rating prediction $\hat{\mathbf{Y}} = \mu \mathbf{1}\mathbf{1}^\intercal + \mathbf{a}\mathbf{1}^\intercal + \mathbf{1}\mathbf{b}^\intercal + \mathbf{A}\mathbf{B}^\intercal$.

Speaker notes

No notes on this slide.

LATENT FACTOR MODEL FOR RECOMMENDATIONS (2 / 2)

- $\mu \sim \text{Normal}\left(0, 100^2\right)$.
- $\mathbf{a} \sim \text{Normal}\left(0, \kappa_a^2\right)$.
- $\mathbf{A} \sim \text{Normal}\left(0, \kappa_A^2\right)$.
- $\mathbf{b} \sim \text{Normal}\left(0, \kappa_b^2\right)$.
- $\mathbf{B} \sim \text{Normal}\left(0, \kappa_B^2\right)$.
- $Y_{ij} \sim \text{Normal}\left(\hat{Y}_{ij}, \sigma^2\right)$ for observed ratings.

Speaker notes

- We use standard priors for the model parameters, e.g., diffuse but proper prior on μ and shrinkage priors on all other model components.
- κ_a captures variability in user satisfaction, e.g., large κ_a corresponds to some users giving very high ratings on average and some very low ratings on average.
- κ_b captures variability in movie quality, e.g., large κ_b means there is variability on movie quality whereas small κ_b means movies receive very similar ratings on average.

SAMPLING IS INTRACTABLE

- Number of parameters is $\gtrsim 10^6$.
- 32m data points—evaluating the likelihood even once is expensive.
- Posterior has complex posterior due to
  - additive degeneracy in $\hat{\mathbf{Y}}$,
  - rotational invariance of inner product $\mathbf{A}\mathbf{B}^\intercal$.

Speaker notes

- Recall from lecture 12 that highly correlated posteriors lead to poor sampler performance.
- Additive degeneracies arise because increasing μ by δ and decreasing all elements of $\mathbf{a}$ by δ leaves the predictions unchanged.
- Even if we could evaluate the likelihood efficiently, sampling would still be very slow.

VARIATIONAL INFERENCE FOR LATENT FACTOR MODEL

We use a “mean-field” approximation

$$q\left(\mu, \mathbf{a}, \ldots; \boldsymbol{\phi}\right) = q_{\mu}\left(\mu; \boldsymbol{\phi}_{\mu}\right) q_{a}\left(\mathbf{a}; \boldsymbol{\phi}_{a}\right) \ldots$$

such that the variational loss L is tractable and optimize using expectation-maximization-style algorithm: each set of parameters $\boldsymbol{\phi}_{\mu}$, $\boldsymbol{\phi}_{a}$, etc. is updated in turn.

Speaker notes

No notes on this slide.

SO WHAT DO WE GAIN (RECOMMENDER SYSTEM)?

- μ: grand mean ratings without much use.
- $\mathbf{a}$: user-specific means of how easily they are pleased. Interesting but maybe not useful.
- $\mathbf{b}$: movie-specific means, representing quality.
- $\mathbf{A}$: user embeddings, indicating similar tastes.
- $\mathbf{B}$: movie embeddings, indicating similar genres, content, styles, etc.
- $\hat{\mathbf{Y}}$: rating point estimates.
- $p\left(\mathbf{Y}_{\text{mis}} \mid \mathbf{Y}_{\text{obs}}\right)$: posterior predictive for missing ratings.

Speaker notes

No notes on this slide.

SO WHAT DO WE GAIN (SINGLE-CELL SEQUENCING, 1 / 2)?

- μ: overall expression level.
- $\mathbf{a}$: cell-specific expression, indicating overall activity.
- $\mathbf{b}$: gene-specific expression, indicating how common a gene is.
- $\mathbf{A}$: cell embeddings, indicating similar expression profiles, e.g., different tissue types, different individuals in mixed data, different pathologies.
- $\mathbf{B}$: gene embeddings, indicating co-expression.

Speaker notes

- We can “search” in the embedding space to find interesting genes. Suppose we can identify a subset of cancerous cells. Then we can rank all the genes by how well they “align” (in the inner product sense) with the embeddings of the cancerous cells to generate hypotheses for genes associated with pathologies.

SO WHAT DO WE GAIN (SINGLE-CELL SEQUENCING, 2 / 2)?

- $\hat{\mathbf{Y}}$: expression point estimates.
- $p\left(\mathbf{Y}_{\text{mis}} \mid \mathbf{Y}_{\text{obs}}\right)$: posterior predictive for unobserved expressions.

Speaker notes

- We can use the posterior predictive distribution for unobserved cell-gene pairs to detect anomalous expressions.

SCALING TO MASSIVE DATASETS

- Datasets are growing rapidly in all domains. Netflix now has 280m subscribers and 100bn hours of content viewed—11.4m years.
- Even fitting these data into memory is challenging.
- How can we use all the data without needing enormous compute clusters—and a lot of money?

Speaker notes

- The number of hours watched on Netflix is almost 0.1% of the age of the universe (13.8bn years).

STOCHASTIC VARIATIONAL INFERENCE (1 / 5)

Suppose observations are conditional independent given parameters, i.e.,

$$p\left(\mathbf{y} \mid \boldsymbol{\theta}\right) = \prod_{i=1}^{n} p\left(y_i \mid \boldsymbol{\theta}\right).$$

Speaker notes

No notes on this slide.

STOCHASTIC VARIATIONAL INFERENCE (2 / 5)

Then the variational loss simplifies to

$$L = \int d\boldsymbol{\theta}\, q(\boldsymbol{\theta}) \left[\log \left(\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} \right) - \sum_{i=1}^{n} \log p(y_i \mid \boldsymbol{\theta}) \right]$$

$$= D\left[q(\boldsymbol{\theta}) \| p(\boldsymbol{\theta}) \right] - \sum_{i=1}^{n} \int d\boldsymbol{\theta}\, q(\boldsymbol{\theta}) \log p(y_i \mid \boldsymbol{\theta}) \,.$$

Speaker notes

No notes on this slide.

STOCHASTIC VARIATIONAL INFERENCE (3 / 5)

- The first term $D\left[q\left(\boldsymbol{\theta}\right)\|p\left(\boldsymbol{\theta}\right)\right]$ is the Kullback-Leibler divergence between approximation and prior. It is independent of sample size.
- The second term $K=\sum_{i=1}^{n}\int d\boldsymbol{\theta}\, q\left(\boldsymbol{\theta}\right)\log p\left(y_i\mid\boldsymbol{\theta}\right)$ is a simple sum over the dataset.

Speaker notes

- Sanity check: If there are no data, we try to approximate the prior by minimizing $D\left[q\left(\boldsymbol{\theta}\right)\|p\left(\boldsymbol{\theta}\right)\right]$.
- Sanity check: If n is large, we maximize an expression that looks a lot like the log likelihood —the data dominate the inference and overwhelm the prior.

STOCHASTIC VARIATIONAL INFERENCE (4 / 5)

We can construct an unbiased estimator

$$\hat{K} = \frac{n}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \int d\boldsymbol{\theta}\, q(\boldsymbol{\theta}) \log p(y_i \mid \boldsymbol{\theta})$$

of K using a subset $\mathcal{B}$ of the data.

Speaker notes

- We take random subsets $\mathcal{B}$ of the data.
- Because K is a sum over i.i.d. samples, $\hat{K}$ is an unbiased estimator.
- The mean and variance grow linearly with the size $|\mathcal{B}|$ of the subset, and thus the relative error scales as $|\mathcal{B}|^{-1/2}$.
- Often using small subsets (also called mini-batches) is sufficient because increasing the batch size has diminishing returns due to the square-root scaling of the relative error.

STOCHASTIC VARIATIONAL INFERENCE (5 / 5)

- We can use $\hat{K}$ on mini-batches of the data instead of K to evaluate gradients and run the optimization.
- The noise from mini-batch sampling gives rise to *stochastic* variational inference.
- The training regime for these approximate posteriors is just like any other deep learning training run.

Speaker notes

No notes on this slide.

VARIATIONAL INFERENCE IN PRACTICE (1 / 2)

- Use a probabilistic programming language like Stan, pyro, or numpyro to do most of the work for you.
- Often the variational loss is not tractable, but we can approximate gradients by sampling from the approximate posterior q similar to the mini-batch loss. This is called black-box variational inference.

Speaker notes

No notes on this slide.

VARIATIONAL INFERENCE IN PRACTICE (2 / 2)

- Like all machine learning training, getting the hyperparameters of the optimizer right and diagnosing convergence is tricky.
- Remember that the variational approximation is … an approximation. It often gets tail behavior wrong. Variational inference is great for scalable prediction but not suitable if the tail probabilities of credible intervals are important.

Speaker notes

- For example, variational inference may be useful for hypothesis generation from large single-cell sequencing data. But it is likely unsuitable for estimating the probability of rare side effects of a drug.