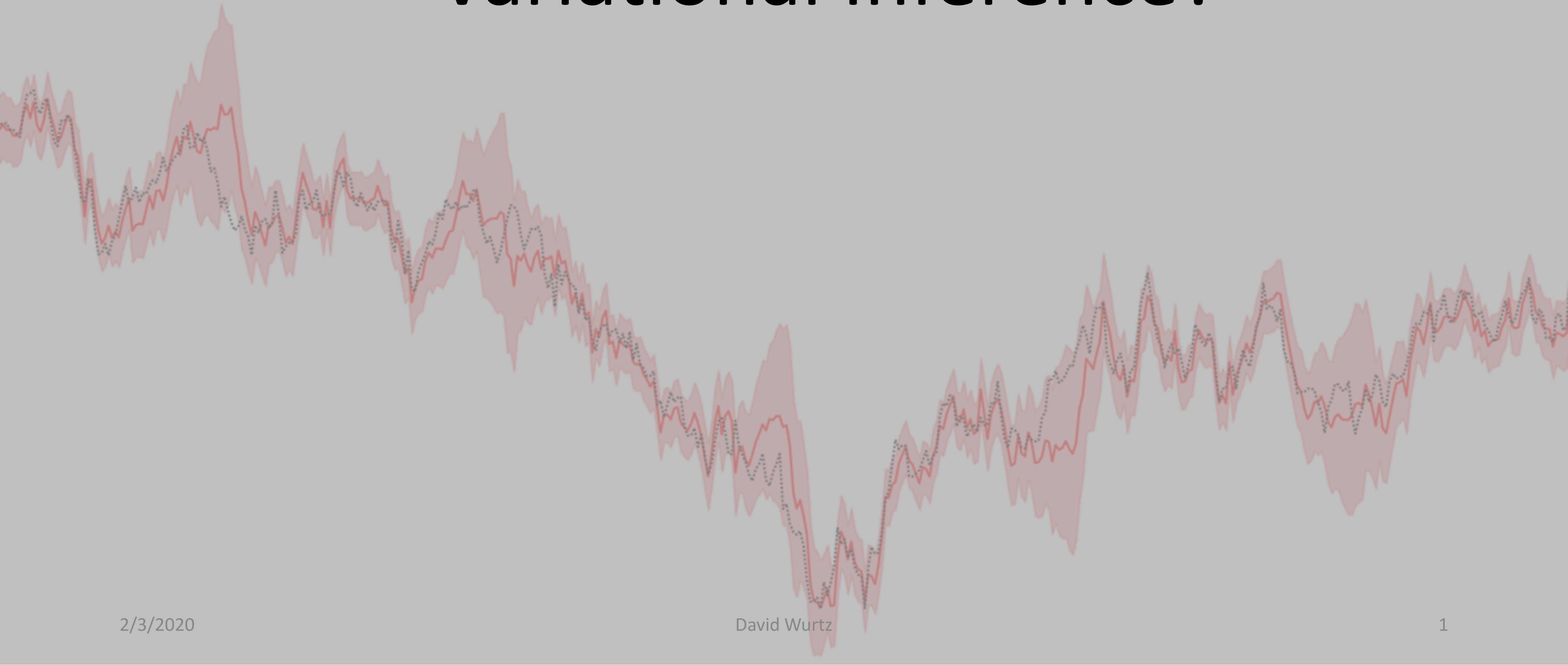
# What’s so cool about Variational Inference?



# Outline

- Motivation
- Background
  - Filtering Posterior
  - Variational Inference
- The cool stuff
  - Variational Bayesian Filtering
  - Modern Variational Inference
- Open Questions
- Conclusion

# Motivation

- Speech processing algorithms often have parameters that are “trained” or “tuned”.
- Often there is no one tuning that works well for a population of users or scenarios.

# Problem Examples

- Smart Mute: earbud bone-conduction is difficult
- Boom-mic beamforming: users move boom around
- Industrial Designers: change the acoustics
- Noise cancelling earbuds: very fit dependent

# Solution

- Algorithms should be “data adaptive” or “self-calibrating”
- Get better as more data arrives

$$P(parameters[t]|data[1:t])$$

# Limitations

- Kalman Filter
  - Both observation and transition densities must Linear and Gaussian
- Unscented Kalman Filter
  - Only gives first and second-order statistics of posterior
- Particle Filter
  - Require lots of particles (MIPS and memory)

# Background: Filtering Posterior

- Filtering posterior and posterior inference in general is very expensive
- Only special cases are tractable
  - Kalman Filter

$$\mu_t, \Sigma_t \leftarrow f(\mu_{t-1}, \Sigma_{t-1}, x_t)$$

# Approximate Filtering Posterior

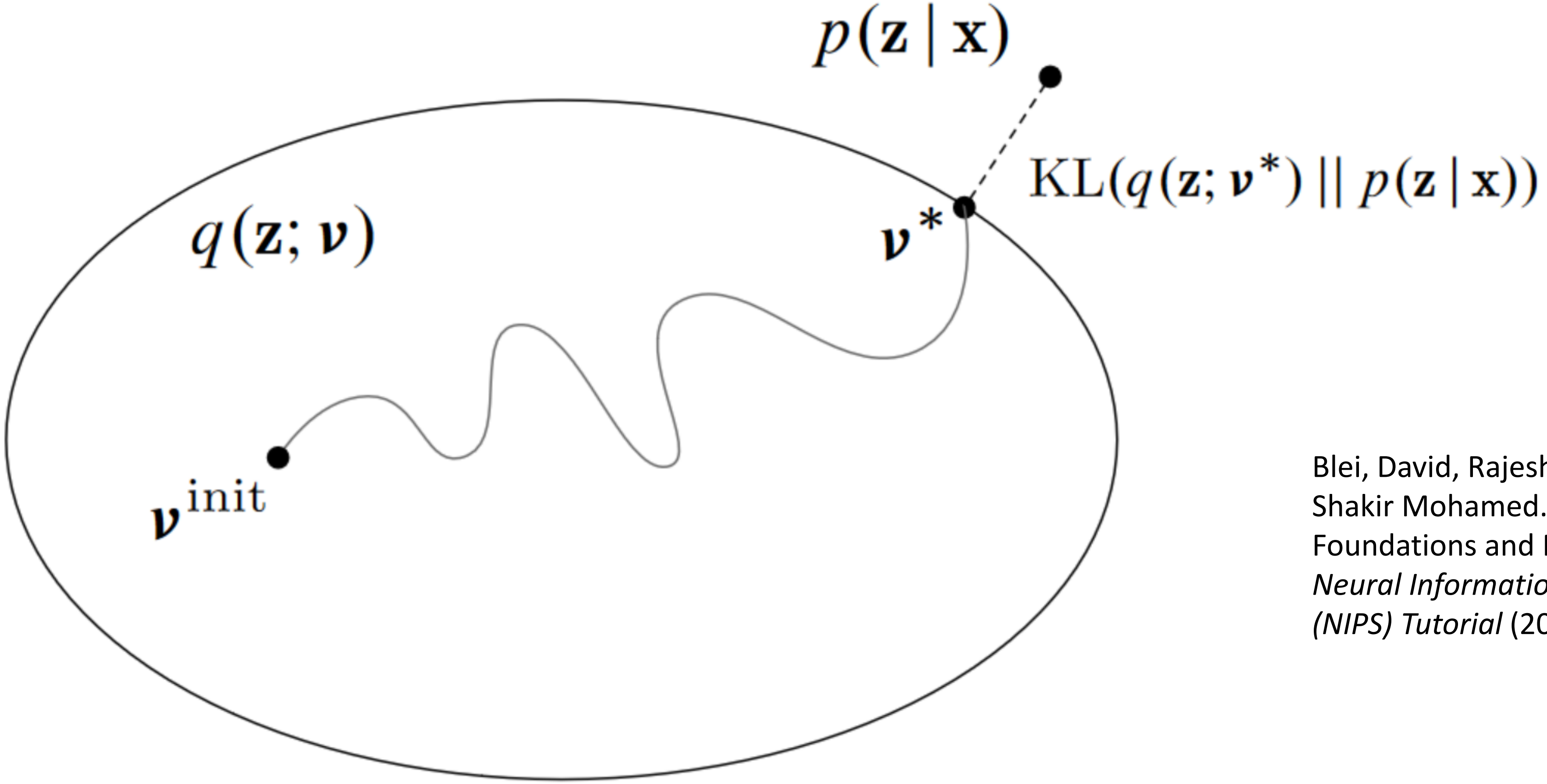
- Unscented Kalman Filter
- Particle Filter

# Background: Variational Inference

- What is “Variational”?

Formulate the thing you want as the solution to an optimization problem.

# Variational Inference




Blei, David, Rajesh Ranganath, and Shakir Mohamed. "Variational Inference: Foundations and Modern Methods." *Neural Information Processing Systems (NIPS) Tutorial* (2016).

# Variational Objective

- Definition of KL:

$$KL(q(z) \parallel p(z|x)) \coloneqq \mathbb{E}_{q(z)}\left[\log \frac{q(z)}{p(z|x)}\right]$$

- Recall from Bayes' rule:

$$p(z|x) = \frac{p(z,x)}{p(x)}$$

- Substituting into KL:

$$KL(q(z) \parallel p(z|x)) = \mathbb{E}_{q(z)}[\log q(z) - \log p(z,x) + \log p(x)]$$

- Simplifying:

$$KL\big(q(z) \parallel p(z|x)\big) = \mathbb{E}_{q(z)}[\log q(z) - \log p(z,x)] + \log p(x)$$

# The ELBO

- Rearranging to have an expression for $\log p(x)$:

$$\log p(x) = \mathbb{E}_{q(z)}[\log p(z,x) - \log q(z)] + KL\big(q(z) \parallel p(z|x)\big)$$

- Maximizing the Evidence Lower BOund necessarily minimizes the KL term. Why?
  - $\log p(x)$ is fixed
  - KL is non-negative

# The Variational Family

- The way that $q(z)$ factorizes is a design choice
- Simplest choice is the *Mean-Field Variational Family*

$$q(z) = \prod_{j=1}^{N} q_j(z_j)$$

# Optimizing the ELBO

- Solutions to the optimization have the following form:

$$q_j^\star(z_j) \propto \exp\left(\mathbb{E}_{q(z_{-j})}\left[\log p(z_j, z_{-j}, x)\right]\right)$$

- For derivations and examples:
  - Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. "Variational inference: A review for statisticians." *Journal of the American statistical Association* 112.518 (2017): 859-877.
  - Bishop, Christopher M. *Pattern recognition and machine learning*. springer, 2006.
  - Murphy, Kevin P. *Machine learning: a probabilistic perspective*. MIT press, 2012.

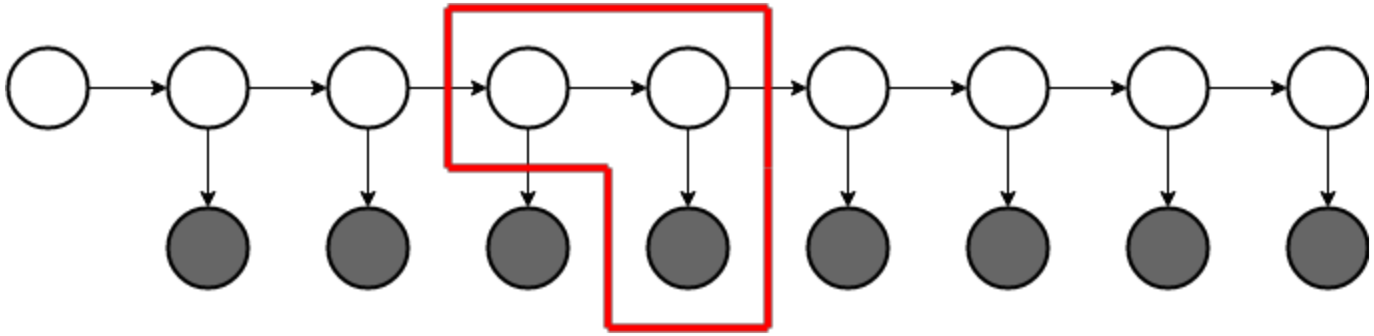
# Limitations

- You must be able to write down the model: $p(z_j, z_{-j}, x)$
- The expectations $\mathbb{E}_{q(z_{-j})}\left[\log p(z_j, z_{-j}, x)\right]$must be analytically tractable
- Derivations are:
  - model-specific
  - fantastically tedious and error-prone (though the result is often elegant)

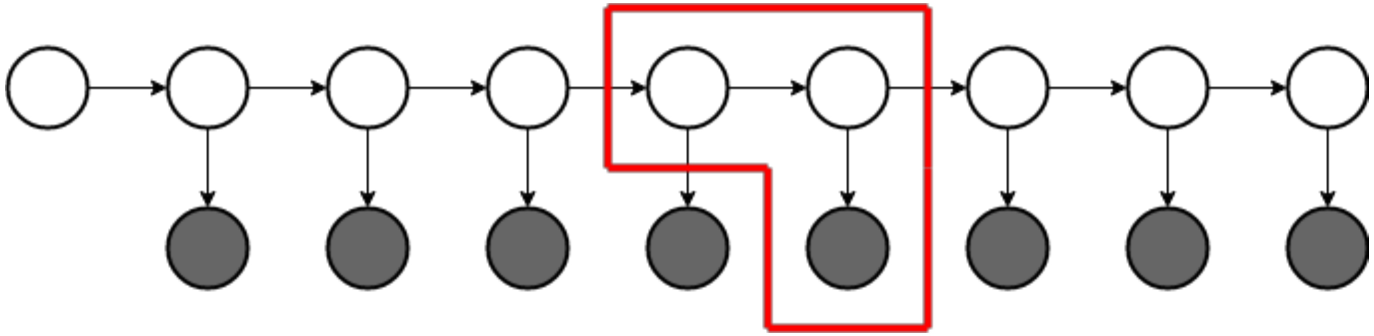
# Recap

- Motivation
- Background
  - Filtering Posterior
  - Variational Inference
- The cool stuff
  - Variational Bayesian Filtering
  - Modern Variational Inference
- Open Questions
- Conclusion

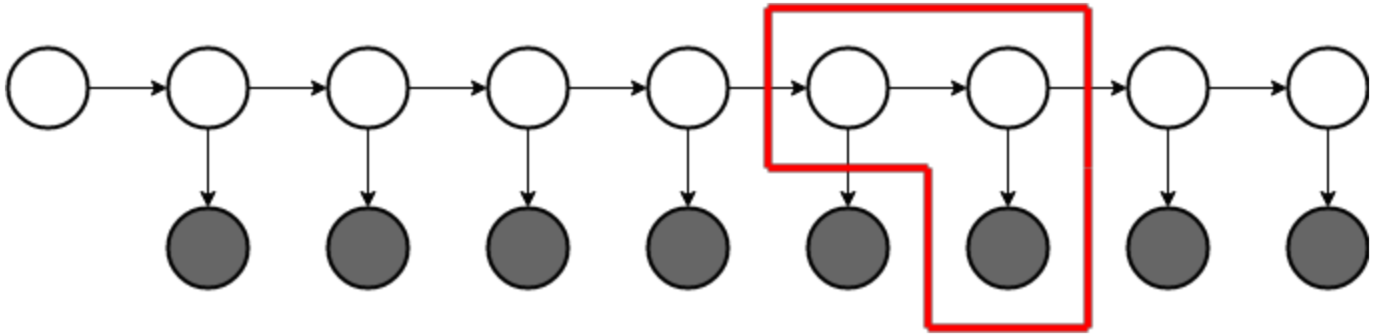
# The cool stuff: Variational Bayesian Filtering



# The cool stuff: Variational Bayesian Filtering



# The cool stuff: Variational Bayesian Filtering



# Bayesian Filtering

- The "local" joint:

$$p(z_t, z_{t-1}, x_t | x_{1:t-1}) = p(x_t | z_t) p(z_t | z_{t-1}) p(z_{t-1} | x_{1:t-1})$$

(chain rule)

# Bayesian Filtering

- The "local" joint:

$$p(z_t, z_{t-1}, x_t | x_{1:t-1}) = p(z_t, z_{t-1} | x_{1:t}) p(x_t | x_{1:t-1})$$

(also chain rule)

# Bayesian Filtering

- Using Bayes' rule:

$$p(z_t, z_{t-1}|x_{1:t}) = \frac{p(x_t|z_t)p(z_t|z_{t-1})p(z_{t-1}|x_{1:t-1})}{p(x_t|x_{1:t-1})}$$

# Bayesian Filtering

- The Filtering Posterior:

$$p(z_t|x_{1:t}) = \int p(z_t, z_{t-1}|x_{1:t})dz_{t-1}$$

(marginalizing out $z_{t-1}$)

# Variational Bayesian Filtering

- The setup needs 2 things…
  - The joint:

$$p(z_t, z_{t-1}, x_t | x_{1:t-1}) = p(x_t|z_t)p(z_t|z_{t-1})q(z_{t-1}|x_{1:t-1})$$

  - Choice of variational family:

$$q(z_t, z_{t-1}|x_{1:t}) = \prod_{j=1}^{N} q_j(z_{j,t}, z_{j,t-1}|x_{1:t})$$

For example if $z = \{\mu, \Sigma\}$, then $q(z_t, z_{t-1}|x_{1:t})$ could factor this way:

$q(\mu_t, \Sigma_t, \mu_{t-1}, \Sigma_{t-1}|x_{1:t}) = q(\mu_t, \mu_{t-1}\,|x_{1:t})\; q(\Sigma_t, \Sigma_{t-1}|x_{1:t})$

# The ELBO

- Same objective as before:

$$\mathbb{E}_{q(z_t, z_{t-1}|x_{1:t})}[\log p(z_t, z_{t-1}, x_t|x_{1:t-1}) - \log q(z_t, z_{t-1}|x_{1:t})]$$

# Optimizing the ELBO

- Same form of solution as before:

$$q_j^\star(z_{j,t}, z_{j,t-1}|x_{1:t}) \propto \exp\left(\mathbb{E}_{q(z_{-j,t}, z_{-j,t-1}|x_{1:t})}[\log p(z_t, z_{t-1}, x_t|x_{1:t-1})]\right)$$

- End up with a recurrent expression for the shaping parameters, $v_t$, of the variational filtering posterior:

$$v_t \leftarrow g(v_{t-1}, x_t)$$

- These look like Kalman Filters:

$$\mu_t, \Sigma_t \leftarrow f(\mu_{t-1}, \Sigma_{t-1}, x_t)$$

# Examples

- A range of easy to complex examples:

  Šmídl, Václav, and Anthony Quinn. *The variational Bayes method in signal processing*. Springer Science & Business Media, 2006.

- A frequency-domain audio processing example:

  S. Malik, J. Benesty, and J. Chen, “A Bayesian framework for blind adaptive beamforming,” *IEEE Tran. on Signal Processing*, vol. 62, no. 9, pp. 2370–2384, 2014.

# Limitations

- Not a lot of tractable variational families for $q(z_t, z_{t-1} | x_{1:t})$
- You must be able to write down $p(z_t, z_{t-1}, x_t | x_{1:t-1})$
- Expectations must be analytically tractable
  - Problem example:

$$\mathbb{E}_{q(\Sigma_1)}[\log \det(\Sigma_1 + \Sigma_2)]$$

# The Cool Stuff: Modern Variational Inference

- Helps address limitations of (Classical) Variational Inference
  - Variational posterior can be learned
  - The model can be learned
  - Variational family can be learned

Blei, David, Rajesh Ranganath, and Shakir Mohamed. "Variational Inference: Foundations and Modern Methods." *Neural Information Processing Systems (NIPS) Tutorial* (2016).

# The Main Ideas

- Intractable expectations are approximated with Monte Carlo
- Unknown parameters are learned
- Unknown functions are approximated (e.g. with a DNN)

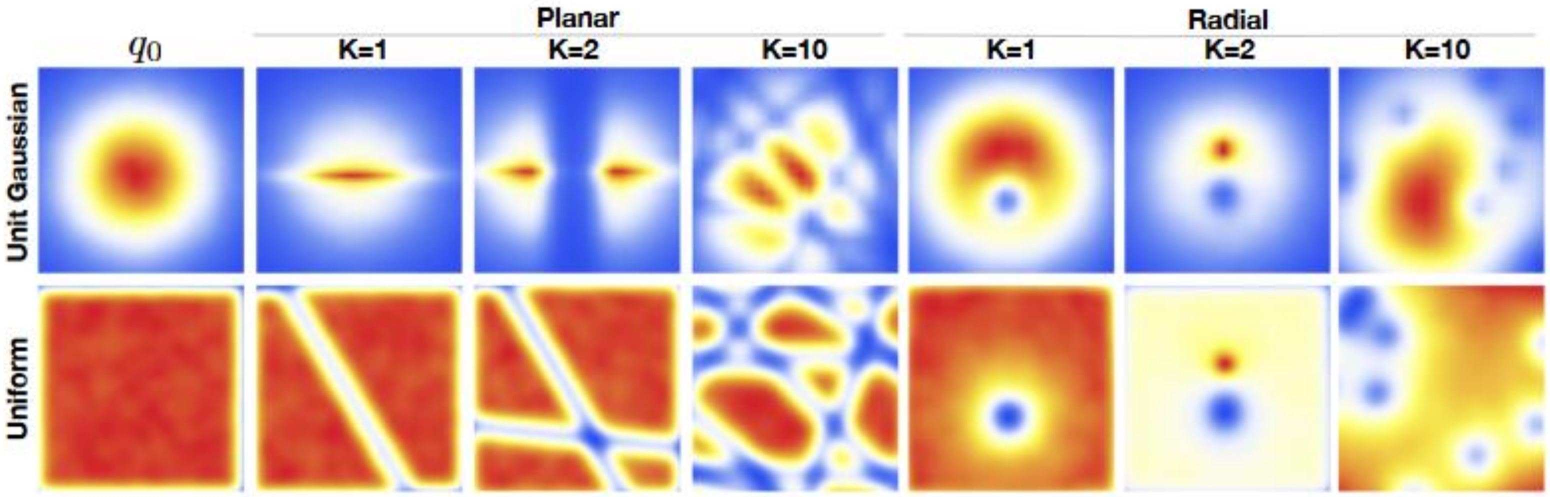
$$\mathbb{E}_{z\sim q(z|x)}\left[\log p_\phi(z,x) - \log q_\theta(z|x)\right]$$

- For example

$$p_\phi(x|z) = Normal\left(\mu_\phi(z), \Sigma_\phi(z)\right)$$
$$q_\theta(z|x) = Normal\left(\mu_\phi(x), \Sigma_\phi(x)\right)$$

# The Main Ideas

- You can learn a complex probability distribution by transforming known simple distribution with a composition of learned invertible functions




Rezende, Danilo Jimenez, and Shakir Mohamed. "Variational inference with normalizing flows." *arXiv preprint arXiv:1505.05770* (2015).

# Open Questions

- The ideas from Modern Variational Inference look like they could be applied to the filtering posterior problem.
  - Given some observations that were generated from known linear gaussian model, how well can you learn a Kalman Filter?
  - How well can Variational Bayes Filters be learned instead of hand-derived?

# Conclusion

- Variational Inference gives us a way derive novel filters for many problems.
  - However, we often need to make compromises in our modeling decisions so that derivations are tractable.

- Modern Variational Inference suggests that these filters can be learned, rather than derived.
  - Allowing us to explore models that otherwise wouldn’t be considered.