Variational Inference

Statistics · Module 9 — Computational Statistics

Node 9.6

Position in Knowledge Graph

System: Statistics

Structural Domain: Computational Statistics

Node: 9.6

Variational inference converts the problem of computing an intractable posterior distribution into an optimization problem. Where MCMC (Node 9.3) produces asymptotically exact samples at high computational cost, VI produces a fast deterministic approximation whose quality depends on the expressiveness of the chosen variational family. This node develops the ELBO, the mean-field approximation, the CAVI algorithm, the directional asymmetry of KL divergence, and the extension to stochastic variational inference for large datasets.

Formal Definition

Let $\pi(\theta \mid x) = p(\theta \mid x)$ be a posterior distribution on $\Theta \subseteq \mathbb{R}^k$ that is intractable – either the normalizing constant $m(x) = \int p(x, \theta)\, d\theta$ cannot be computed or MCMC mixing is prohibitively slow.

Let $\mathcal{Q}$ be a family of tractable distributions on Θ. **Variational inference** seeks

$$q^* = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}(q(\theta) \| p(\theta \mid x)).$$

Since $\mathrm{KL}(q \| p(\cdot \mid x))$ involves the unknown normalizing constant $m(x)$, VI instead maximizes the **Evidence Lower BOund (ELBO)**:

$$\mathcal{L}(q) = \mathbb{E}_q[\log p(x, \theta)] - \mathbb{E}_q[\log q(\theta)] = \mathbb{E}_q[\log p(x, \theta)] + H(q),$$

where $H(q) = -\mathbb{E}_q[\log q(\theta)]$ is the entropy of q.

Intuition

Bayesian inference requires integrating over the entire parameter space. MCMC does this by simulating a random walk and averaging. Variational inference replaces the integral with a search: find the distribution in a tractable family that is closest to the true posterior. "Closest" is measured by KL divergence, and the ELBO provides a computable objective that avoids the normalizing constant. The approximation quality depends entirely on whether the variational family $\mathcal{Q}$ can capture the shape of the true posterior – if the posterior is multimodal and $\mathcal{Q}$ contains only unimodal distributions, the approximation will miss important structure.

Structural Role

Variational inference trades exactness for scalability. It depends on the ELBO framework established in Node 9.5 (EM) and serves as the scalable alternative to MCMC (Node 9.3) for large datasets and complex models. EM is the special case where the variational family is unrestricted (the E-step uses the exact posterior over latent variables). Variational inference is the foundation of variational autoencoders (VAEs) and amortized inference in modern machine learning.

Mathematical Development

The ELBO Decomposition

For any $q(\theta)$,

$$\log m(x) = \mathcal{L}(q) + \mathrm{KL}(q(\theta) \| p(\theta \mid x)).$$

Proof. Expanding the KL divergence:

$$\mathrm{KL}(q \| p(\cdot \mid x)) = \mathbb{E}_q\left[\log \frac{q(\theta)}{p(\theta \mid x)}\right] = \mathbb{E}_q[\log q(\theta)] - \mathbb{E}_q[\log p(\theta \mid x)].$$

Since $p(\theta \mid x) = p(x, \theta)/m(x)$:

$$\mathrm{KL}(q \| p(\cdot \mid x)) = \mathbb{E}_q[\log q(\theta)] - \mathbb{E}_q[\log p(x, \theta)] + \log m(x).$$

Rearranging:

$$\log m(x) = \underbrace{\mathbb{E}_q[\log p(x, \theta)] - \mathbb{E}_q[\log q(\theta)]}_{\mathcal{L}(q)} + \mathrm{KL}(q \| p(\cdot \mid x)). \quad \square$$

Since $\mathrm{KL} \geq 0$, $\mathcal{L}(q) \leq \log m(x)$, with equality iff $q = p(\theta \mid x)$. Thus maximizing the ELBO over $\mathcal{Q}$ is equivalent to minimizing $\mathrm{KL}(q \| p(\cdot \mid x))$ over $\mathcal{Q}$.

Alternative Form of the ELBO

The ELBO can be decomposed as

$$\mathcal{L}(q) = \mathbb{E}_q[\log p(x \mid \theta)] - \mathrm{KL}(q(\theta) \| p(\theta)),$$

where $p(\theta)$ is the prior. This form reveals the tradeoff: the first term rewards q for placing mass where the likelihood is large; the second term penalizes q for deviating from the prior.

Mean-Field Approximation

The **mean-field family** assumes full factorization:

$$\mathcal{Q}_{\mathrm{MF}} = \left\{ q(\theta) = \prod_{j=1}^{k} q_j(\theta_j) \right\}.$$

Theorem (Optimal Mean-Field Update). Fixing all factors q_i for $i \neq j$, the optimal q_j satisfies

$$\log q_j^*(\theta_j) = \mathbb{E}_{q_{-j}}[\log p(x, \theta)] + \text{const},$$

where $\mathbb{E}_{q_{-j}}$ denotes expectation over all components except j, and the constant ensures q_j^* normalizes to 1.

Proof. Write the ELBO as

$$\mathcal{L}(q) = \int q_j(\theta_j) \left[\int \prod_{i \neq j} q_i(\theta_i) \log p(x, \theta) \, d\theta_{-j} \right] d\theta_j - \int q_j(\theta_j) \log q_j(\theta_j) \, d\theta_j + C,$$

where C collects terms not depending on q_j. Define $f_j(\theta_j) = \mathbb{E}_{q_{-j}}[\log p(x, \theta)]$. Then

$$\mathcal{L}(q) = \int q_j(\theta_j) f_j(\theta_j) \, d\theta_j - \int q_j(\theta_j) \log q_j(\theta_j) \, d\theta_j + C.$$

This is maximized when $q_j(\theta_j) \propto \exp(f_j(\theta_j))$, since the functional is $-\text{KL}(q_j \| \tilde{q}_j) + \text{const}$ where $\tilde{q}_j \propto \exp(f_j)$. □

Coordinate Ascent Variational Inference (CAVI)

CAVI cycles through coordinates $j = 1, \ldots, k$, updating each q_j according to the optimal update while holding q_{-j} fixed. Each update increases (or maintains) the ELBO, guaranteeing convergence to a local optimum. The algorithm terminates when the relative change in ELBO falls below a threshold.

For exponential family models with conjugate priors, each CAVI update yields a distribution in the same exponential family, and the update reduces to updating the natural parameters:

$$\eta_j^{(t+1)} = \eta_j^{\text{prior}} + \mathbb{E}_{q_{-j}^{(t)}}[t_j(x, \theta_{-j})],$$

where η_j^{prior} are the prior natural parameters and t_j are the sufficient statistics contributed by factor j.

KL Divergence Asymmetry

The choice of KL direction has profound consequences:

$\text{KL}(q\|p)$ (variational inference): This is **mode-seeking** (or zero-avoiding). Where $p(\theta \mid x) \approx 0$, q must also be approximately 0 (otherwise the KL divergence diverges). Consequently, q tends to concentrate on a single mode of p and underestimates posterior variance.

$\text{KL}(p\|q)$ (expectation propagation): This is **mean-seeking** (or mass-covering). Where $p(\theta \mid x) > 0$, q must also be positive. Consequently, q spreads to cover all modes, typically overestimating variance.

For a bimodal posterior with modes at $\pm\mu$: - $\text{KL}(q\|p)$-optimal Gaussian: concentrates on one mode, $q^* \approx \mathcal{N}(\mu, \sigma^2)$ or $\mathcal{N}(-\mu, \sigma^2)$. - $\text{KL}(p\|q)$-optimal Gaussian: centers at 0 with large variance, $q^* \approx \mathcal{N}(0, \mu^2 + \sigma^2)$.

Relation to EM

EM (Node 9.5) is variational inference with $\mathcal{Q} = \{p(Z \mid Y, \theta) : \theta \in \Theta\}$, the family of exact posteriors over latent variables parameterized by θ. The E-step sets $q(Z) = p(Z \mid Y, \theta^{(t)})$ (exact posterior, KL gap = 0), and the M-step optimizes θ. Variational EM replaces the exact E-step with an approximate one: $q(Z) \in \mathcal{Q}_{\text{approx}}$, enabling inference when the E-step is itself intractable.

Stochastic Variational Inference

For large datasets $x = (x_1, \ldots, x_n)$, CAVI requires a full pass through the data at each update. **Stochastic variational inference (SVI)** exploits the fact that the ELBO decomposes as a sum over data points (for conditionally i.i.d. models):

$$\mathcal{L}(q) = \sum_{i=1}^{n} \mathbb{E}_q[\log p(x_i \mid \theta)] - \mathrm{KL}(q(\theta) \| p(\theta)).$$

SVI replaces the full-data gradient with a noisy mini-batch estimate:

$$\hat{\nabla}\mathcal{L} = \frac{n}{|S|} \sum_{i \in S} \nabla_\lambda \mathbb{E}_q[\log p(x_i \mid \theta)] - \nabla_\lambda \mathrm{KL}(q_\lambda \| p),$$

where S is a random mini-batch and λ parameterizes q. With a Robbins–Monro step size schedule ($\sum \alpha_t = \infty$, $\sum \alpha_t^2 < \infty$), SVI converges to a local optimum of the ELBO.

Worked Examples

Example 1: Mean-Field VI for Normal-Normal Model

Let $x_1, \ldots, x_n \mid \mu, \tau \sim \mathcal{N}(\mu, \tau^{-1})$ with priors $\mu \sim \mathcal{N}(0, \sigma_0^2)$ and $\tau \sim \mathrm{Gamma}(a_0, b_0)$. The mean-field approximation is $q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau)$.

Update for q_μ:

$$\log q_\mu^*(\mu) = \mathbb{E}_{q_\tau}\left[-\frac{\mathbb{E}[\tau]}{2} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{\mu^2}{2\sigma_0^2} \right] + \text{const.}$$

This is a quadratic in μ, so $q_\mu^* = \mathcal{N}(\mu_q, \sigma_q^2)$ with

$$\sigma_q^2 = (n\mathbb{E}_{q_\tau}[\tau] + 1/\sigma_0^2)^{-1}, \quad \mu_q = \sigma_q^2 \cdot n\mathbb{E}_{q_\tau}[\tau]\bar{x}.$$

Update for q_τ:

$$q_\tau^* = \mathrm{Gamma}\left(a_0 + n/2,\ b_0 + \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}_{q_\mu}[(x_i - \mu)^2] \right).$$

CAVI alternates these updates until the ELBO converges. With $n = 100$, $\bar{x} = 2.1$, convergence is typically achieved in 10–20 iterations. The mean-field approximation underestimates the posterior covariance between μ and τ (which is zero in the factorized approximation but nonzero in the true posterior).

Example 2: KL Direction Comparison

Consider a posterior $p(\theta \mid x) = 0.5\,\mathcal{N}(-3, 0.5^2) + 0.5\,\mathcal{N}(3, 0.5^2)$ and restrict $\mathcal{Q}$ to Gaussians $q = \mathcal{N}(m, s^2)$.

Minimizing $\mathrm{KL}(q\|p)$: the optimizer selects one mode, yielding $q^* \approx \mathcal{N}(3, 0.5^2)$ (or symmetrically $\mathcal{N}(-3, 0.5^2)$). The KL divergence penalizes placing mass where p is near zero (the region between modes), so q collapses to a single mode.

Minimizing $\mathrm{KL}(p\|q)$: the optimizer must cover both modes, yielding $q^* \approx \mathcal{N}(0, 9.25)$. This captures both modes but vastly overestimates the spread, assigning substantial mass to the region between modes where p is nearly zero.

Cross-Links

- **Linear Algebra:** The CAVI updates for Gaussian mean-field approximations reduce to solving linear systems; the natural gradient of the ELBO with respect to the natural parameters of an exponential family q is the Euclidean gradient in the natural parameterization.
- **AI:** Variational autoencoders (VAEs) use amortized variational inference: an encoder network maps each data point x_i to variational parameters λ_i, optimizing the ELBO jointly over the encoder and decoder. The reparameterization trick enables backpropagation through the sampling step.
- **Economics:** Variational Bayes is increasingly used in macroeconometrics for approximating posterior distributions of high-dimensional VAR models where MCMC is too slow for real-time forecasting.
- **Quantum:** Variational quantum algorithms (VQE, QAOA) optimize a parameterized quantum circuit to minimize an energy functional, mirroring the VI principle of optimizing over a tractable family to approximate an intractable distribution.

Structural Compression

Invariant: $\log m(x) = \mathcal{L}(q) + \mathrm{KL}(q \| p(\theta \mid x))$: maximizing the ELBO over $\mathcal{Q}$ minimizes the KL divergence to the posterior, with the gap measuring the approximation quality.

Mechanism: Jensen's inequality applied to the log marginal likelihood yields the ELBO as a lower bound. The mean-field factorization restricts $\mathcal{Q}$ to product distributions, enabling coordinate-wise closed-form updates (CAVI). Each CAVI step increases the ELBO, converging to a local optimum. The mode-seeking direction $\mathrm{KL}(q\|p)$ produces compact approximations that may miss posterior mass, trading coverage for computational tractability.

Cross-System Echo: The ELBO decomposition is the master identity of approximate inference. In EM, the KL gap is zero by construction (exact E-step). In VI, the gap is nonzero but controlled. In VAEs, the ELBO is the training objective, with the encoder parameterizing q and the decoder parameterizing $p(x \mid \theta)$. In information theory, the ELBO is the negative description length under code q; minimizing KL is equivalent to minimizing the excess bits required to describe the data. The pattern – bounding an intractable quantity by a tractable one and optimizing the bound – is the universal computational strategy for problems where exact solutions are infeasible.

Exercises

1. Derive the ELBO decomposition $\log m(x) = \mathcal{L}(q) + \mathrm{KL}(q \| p(\theta \mid x))$ and verify that the ELBO equals the log marginal likelihood when $q = p(\theta \mid x)$.
2. For the normal-normal model with known variance, derive the mean-field CAVI updates for q_μ and q_τ and show that each update is a closed-form exponential family distribution.
3. Prove that CAVI monotonically increases the ELBO at each coordinate update. State precisely what regularity conditions are needed.
4. For a bimodal target $p = 0.5\mathcal{N}(-\mu, 1) + 0.5\mathcal{N}(\mu, 1)$, compute the $\mathrm{KL}(q\|p)$-optimal and $\mathrm{KL}(p\|q)$-optimal Gaussian approximations analytically as functions of μ. Plot both as μ ranges from 0 to 5.
5. Show that EM is a special case of variational inference where the E-step uses the unrestricted family $\mathcal{Q} = \{q(Z) : q \text{ is a density}\}$, so the KL gap is zero after the E-step.
6. For a model with $n = 10{,}000$ observations, estimate the per-iteration cost of CAVI versus SVI with mini-batch size 100. Explain the tradeoff between noise in the gradient and computational savings.
7. Argue, structurally, why the ELBO is the unifying object across EM, variational inference, and variational autoencoders: in each case, the same decomposition $\log p(x) = \mathrm{ELBO} + \mathrm{KL}$ determines the objective,

the approximation gap, and the algorithm, with the only difference being the expressiveness of the variational family and whether the parameters θ are integrated out or optimized.

Statistics · Module 9 — Computational Statistics · Node 9.6
Concepts: variational-principle, shannon-entropy, prior-posterior, convergence
Prerequisites: 9.5, 9.3, 6.1

hbar.university — own.your.cognition