Numerical Optimization

Statistics · Module 9 — Computational Statistics

Node 9.1

Position in Knowledge Graph

System: Statistics

Structural Domain: Computational Statistics

Node: 9.1

Maximum likelihood estimation, M-estimation, and posterior mode computation all require maximizing an objective function $Q(\theta)$ over $\theta \in \Theta \subseteq \mathbb{R}^k$. Closed-form solutions exist only in special cases (exponential families with canonical sufficient statistics). This node develops the principal iterative algorithms for numerical optimization in statistics, their convergence properties, and the structural role of curvature information.

Formal Definition

Let $Q : \Theta \to \mathbb{R}$ be a twice continuously differentiable objective function. A point $\theta^* \in \Theta$ is a **stationary point** if $\nabla Q(\theta^*) = 0$. It is a **local maximizer** if, additionally, the Hessian $\nabla^2 Q(\theta^*)$ is negative definite.

An **iterative optimization algorithm** generates a sequence $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots$ such that $Q(\theta^{(t+1)}) \geq Q(\theta^{(t)})$ (ascent property) and $\theta^{(t)} \to \theta^*$ under appropriate conditions.

The **gradient ascent** update is

$$\theta^{(t+1)} = \theta^{(t)} + \alpha_t \nabla Q(\theta^{(t)}),$$

where $\alpha_t > 0$ is the step size. The **Newton–Raphson** update is

$$\theta^{(t+1)} = \theta^{(t)} - \left[\nabla^2 Q(\theta^{(t)})\right]^{-1} \nabla Q(\theta^{(t)}).$$

Fisher scoring replaces the Hessian with its expectation under the model, yielding

$$\theta^{(t+1)} = \theta^{(t)} + \mathcal{I}(\theta^{(t)})^{-1} \nabla \ell(\theta^{(t)}),$$

where $\mathcal{I}(\theta) = \mathbb{E}_\theta[-\nabla^2 \ell(\theta)]$ is the expected Fisher information.

Intuition

Gradient ascent moves uphill along the steepest direction but treats all directions equally, ignoring the curvature of the objective surface. Newton–Raphson fits a local quadratic approximation and jumps directly to its maximum, which is why it converges much faster near the optimum. The price is computing and inverting the Hessian matrix at each step, which is $O(k^3)$ in parameter dimension.

Fisher scoring is the statistically natural variant: it replaces the observed curvature (which fluctuates with the data) by the expected curvature (which depends only on the model), producing updates that are equivariant under reparameterization. For generalized linear models, Fisher scoring reduces to iteratively reweighted least squares (IRLS).

Structural Role

Numerical optimization is the computational substrate of all likelihood-based and M-estimation inference. Newton–Raphson is the canonical algorithm for computing the MLE; Fisher scoring is its information-geometric counterpart. The convergence rates of these algorithms are controlled by the curvature of the log-likelihood, which is itself the Fisher information (Node 2.5). This node is prerequisite to the EM algorithm (Node 9.5), which replaces a single intractable optimization with a sequence of tractable ones.

Mathematical Development

Convergence of Gradient Ascent

Suppose Q is L-smooth (i.e., $\|\nabla^2 Q(\theta)\| \le L$ for all θ) and m-strongly concave (i.e., $\nabla^2 Q(\theta) \preceq -mI$ for all θ, with $m > 0$). With fixed step size $\alpha = 1/L$,

$$\|\theta^{(t+1)} - \theta^*\| \le \left(1 - \frac{m}{L}\right) \|\theta^{(t)} - \theta^*\|.$$

This is **linear (geometric) convergence** with rate $\kappa = 1 - m/L$, where L/m is the **condition number** of the problem. Poor conditioning ($L \gg m$) produces slow convergence.

Line Search

The **Armijo backtracking** line search selects α_t as the largest value in $\{\beta^0, \beta^1, \beta^2, \dots\}$ (for shrinkage factor $0 < \beta < 1$) satisfying

$$Q(\theta^{(t)} + \alpha_t d^{(t)}) \ge Q(\theta^{(t)}) + c\,\alpha_t \nabla Q(\theta^{(t)})^\top d^{(t)},$$

where $d^{(t)}$ is the search direction and $0 < c < 1$. This guarantees sufficient increase at each step.

Newton–Raphson Convergence

Theorem (Local Quadratic Convergence). Let Q be three times continuously differentiable in a neighborhood of θ^* with $\nabla^2 Q(\theta^*)$ nonsingular. Then there exists $\delta > 0$ such that for $\|\theta^{(0)} - \theta^*\| < \delta$,

$$\|\theta^{(t+1)} - \theta^*\| \le C \|\theta^{(t)} - \theta^*\|^2,$$

where $C = \frac{1}{2}\|\nabla^2 Q(\theta^*)^{-1}\| \sup_\theta \|\nabla^3 Q(\theta)\|$.

Proof sketch. Taylor-expand $\nabla Q(\theta^{(t)})$ around θ^*:

$$\nabla Q(\theta^{(t)}) = \nabla^2 Q(\theta^*)(\theta^{(t)} - \theta^*) + O(\|\theta^{(t)} - \theta^*\|^2).$$

The Newton step gives $\theta^{(t+1)} - \theta^* = \theta^{(t)} - \theta^* - [\nabla^2 Q(\theta^{(t)})]^{-1}\nabla Q(\theta^{(t)})$. Substituting the Taylor expansion and using continuity of $\nabla^2 Q$, the linear term cancels, leaving only the quadratic remainder. □

Trust Regions

When Newton–Raphson is initialized far from θ^*, the Hessian may not be negative definite and the Newton step may overshoot. A **trust region** method restricts the step to a ball:

$$\theta^{(t+1)} = \arg\max_{\|\delta\| \le \Delta_t} \left[\nabla Q(\theta^{(t)})^\top \delta + \frac{1}{2}\delta^\top \nabla^2 Q(\theta^{(t)})\delta \right],$$

where the trust radius Δ_t is adapted based on how well the quadratic model predicts the actual change in Q.

Profile Likelihood

For a partitioned parameter $\theta = (\psi, \lambda)$ where λ is a nuisance parameter, the **profile log-likelihood** is

$$\ell_p(\psi) = \max_{\lambda} \ell(\psi, \lambda) = \ell(\psi, \hat{\lambda}_\psi),$$

where $\hat{\lambda}_\psi$ is the MLE of λ at fixed ψ. Optimization of ℓ_p over ψ yields the marginal MLE, reducing dimension at the cost of iterative inner optimization. The curvature of ℓ_p provides adjusted standard errors for ψ.

Quasi-Newton Methods

When the Hessian is expensive to compute, **quasi-Newton methods** build an approximation $B_t \approx -\nabla^2 Q(\theta^{(t)})$ from successive gradient evaluations. The **BFGS update** maintains a positive definite approximation:

$$B_{t+1} = B_t + \frac{y_t y_t^\top}{y_t^\top s_t} - \frac{B_t s_t s_t^\top B_t}{s_t^\top B_t s_t},$$

where $s_t = \theta^{(t+1)} - \theta^{(t)}$ and $y_t = \nabla Q(\theta^{(t+1)}) - \nabla Q(\theta^{(t)})$. BFGS achieves **superlinear convergence** – faster than linear gradient descent but slower than quadratic Newton – while requiring only gradient evaluations. The limited-memory variant (L-BFGS) stores only the most recent m pairs (s_t, y_t), making it feasible for high-dimensional problems.

Convergence Criteria

Standard termination criteria include:

1. **Gradient norm:** $\|\nabla Q(\theta^{(t)})\| < \varepsilon_g$.
2. **Relative change:** $|Q(\theta^{(t+1)}) - Q(\theta^{(t)})| / (|Q(\theta^{(t)})| + 1) < \varepsilon_f$.
3. **Parameter change:** $\|\theta^{(t+1)} - \theta^{(t)}\| < \varepsilon_\theta$.

In practice, all three are monitored simultaneously. For MLE computation, typical tolerances are $\varepsilon \sim 10^{-8}$.

Worked Examples

Example 1: Newton–Raphson for Gamma MLE

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Gamma}(\alpha, \beta)$ with density $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$. The score equations are

$$\frac{\partial \ell}{\partial \alpha} = n \log \beta - n\psi(\alpha) + \sum_{i=1}^{n} \log x_i = 0, \quad \frac{\partial \ell}{\partial \beta} = \frac{n\alpha}{\beta} - \sum_{i=1}^{n} x_i = 0,$$

where $\psi(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$ is the digamma function. From the second equation, $\hat{\beta} = n\alpha / \sum x_i$. Substituting into the first gives a single equation in α that requires Newton–Raphson iteration:

$$\alpha^{(t+1)} = \alpha^{(t)} - \frac{n \log(\alpha^{(t)}/\bar{x}) - n\psi(\alpha^{(t)}) + \sum \log x_i}{n/\alpha^{(t)} - n\psi'(\alpha^{(t)})}.$$

For data with $n = 50$, $\bar{x} = 3.2$, $\sum \log x_i = 48.7$, initializing at $\alpha^{(0)} = (\bar{x})^2/s^2$ (method of moments), convergence to six decimal places is typical in 4–5 iterations.

Example 2: Fisher Scoring in Logistic Regression

For logistic regression with $P(Y_i = 1 \mid x_i) = \mu_i(\beta) = (1 + e^{-x_i^\top \beta})^{-1}$, the score is

$$\nabla \ell(\beta) = X^\top (y - \mu(\beta)),$$

and the expected Fisher information is

$$\mathcal{I}(\beta) = X^\top W(\beta) X, \quad W = \text{diag}(\mu_i(1 - \mu_i)).$$

Fisher scoring (equivalently, IRLS) iterates

$$\beta^{(t+1)} = (X^\top W^{(t)} X)^{-1} X^\top W^{(t)} z^{(t)},$$

where the working response is $z^{(t)} = X\beta^{(t)} + (W^{(t)})^{-1}(y - \mu^{(t)})$. For a dataset with $n = 100$, $p = 3$, convergence is typically achieved in 5–8 iterations.

Cross-Links

- **Linear Algebra:** Newton–Raphson requires solving a $k \times k$ linear system at each step; the condition number of the Hessian controls convergence rate.
- **AI:** Adam and other adaptive optimizers approximate second-order information using diagonal or low-rank Hessian estimates; natural gradient descent uses the Fisher information matrix as the Riemannian metric.
- **Economics:** Structural estimation in econometrics relies on Newton–Raphson for GMM and maximum likelihood of DSGE models where closed-form solutions are unavailable.
- **Quantum:** Variational quantum eigensolvers use gradient descent on a parameterized quantum circuit; the quantum Fisher information matrix governs the geometry of the parameter space.

Structural Compression

Invariant: The MLE is a fixed point of the score equation $\nabla \ell(\theta) = 0$; Newton–Raphson converges quadratically to this fixed point when initialized in a sufficiently small neighborhood.

Mechanism: Second-order Taylor expansion of Q around the current iterate; the Newton step solves the resulting quadratic approximation exactly. The Hessian encodes local curvature, collapsing the iterative process to a single step for exactly quadratic objectives.

Cross-System Echo: Newton–Raphson is the natural gradient method in the information geometry of the statistical model, where the Fisher information matrix defines the Riemannian metric on the parameter manifold. The tension between first-order methods (cheap but slow) and second-order methods (expensive but fast) recurs in every optimization discipline: gradient descent versus Newton in statistics, SGD versus K-FAC in deep learning, steepest descent versus conjugate gradients in numerical linear algebra.

Exercises

1. Derive the Fisher scoring update for a Poisson regression model with canonical log link and verify that it reduces to an IRLS iteration.
2. For the normal location model $X_i \sim \mathcal{N}(\mu, 1)$, show that Newton–Raphson converges in exactly one step from any initialization. Explain why.
3. Prove that gradient ascent with step size $\alpha = 1/L$ on an L-smooth, m-strongly concave function satisfies $Q(\theta^*) - Q(\theta^{(t)}) \leq (L/2)(1 - m/L)^t \|\theta^{(0)} - \theta^*\|^2$.
4. Consider the Cauchy location model $X_i \sim \text{Cauchy}(\theta, 1)$. Write the Newton–Raphson update for the MLE. Explain why convergence can fail without a good initialization.
5. Show that for an exponential family in canonical form, the observed and expected Fisher information coincide, so Newton–Raphson and Fisher scoring produce identical iterates.
6. Implement backtracking line search for the log-likelihood of a gamma model. Describe how the step size adapts across iterations as the algorithm approaches the MLE.
7. Argue, structurally, why the condition number L/m of the Hessian plays the same role in optimization convergence that the variance ratio plays in statistical efficiency: both quantify how much information the available data (or curvature) provides per unit of computational (or sampling) effort.

Statistics · Module 9 — Computational Statistics · Node 9.1
Concepts: gradient-descent, convergence, convexity, fisher-information
Prerequisites: 7.5, 2.3
Enables: 9.5

hbar.university — own.your.cognition