ADAPTIVE ANNEALING

Andrew Barron and Xi "Rossi" Luo

Yale University, Department of Statistics

Allerton Conference Presentation, September 27, 2007

Outline

- Sampling and Optimization
  - Markov Chain Methods
  - Statistical Motivations
  - Target Distributions
- Types of Markov Chain Samplers
  - Time-homogeneous transitions designed for a fixed target distribution
  - Time-inhomogeneous transitions for evolving sequence of distributions
- Approximate Diffusions
  - Simulated annealing and its shortcoming
  - Adaptive annealing
  - Drift modifier dynamics
- Variable Augmentation
  - State space enlargement for approximate decoupling of equations
  - Adaptive Gibbs annealing
- Optimization of Superpositions of Ridge Functions

Sampling and Optimization

- $L(w)$ is a smooth objective function on R^d

  A bounded function

  At least two bounded derivatives with respect to any of its coordinates

  May have multiple peaks

- Optimization:

  To within a constant factor, maximize $L(w)$ subject to $||w||$ constraint

  Or maximize the Lagrangian $L(w) - \lambda||w||^2$

- Sampling:

  Draw w according to a given density $p(w)$

  Usually the result of several steps of a Markov chain

  Target density $p_\gamma(w)$ with gain γ is given by

$$p_\gamma(w) = \frac{p_0(w)\exp\{\gamma L(w)\}}{c_\gamma}$$

  Gaussian reference or initial density $p_0(w)$

How big should be γ?

- What is a sufficient gain γ for a random draw to have $L(w)$ nearly maximal?
- Assume a bound on the maximum of the norm of the gradient $\nabla L(w)$
- Let $L^* = L(w^*)$ be a (maximal or nearly maximal) value of the objective achieved by a point with finite $||w^*||$
- Lemma: The mean of $L(w)$ for a random w drawn from $p_\gamma(w)$ is at least

$$L^* - \frac{d}{\gamma} \log \frac{A\gamma}{d}$$

  where A depends on $||w^*||$ and on the bound on the gradient.
- For positive L^*, choosing γ at least d times a log-factor is sufficient for the mean of $L(w)$ to be at least $\frac{1}{2}L^*$.

Markov Chain Samplers

- Target

  Want w distributed according to a specified $\pi(w)$ such as $p_\gamma(w)$

- Markov Chain Monte Carlo Methods

  Generate $w_0, w_1, \ldots, w_T$

  Typically use time-homogeneous transition densities:

$$p(\tilde{w}|w)$$

  Invariance: Transition chosen such that it takes $w \sim \pi$ into $\tilde{w} \sim \pi$.

  Initial density $\pi_0(w)$, not equal to the target $\pi(w)$

  Convergence to $\pi(w)$: Can be arbitrarily slow.

- Gibbs Samplers
- Metropolis-Hastings
- Approximate Diffusions

Approximate Diffusions

- Simplest formulation: a time-homogeneous transition with target $\pi(w)$
- Gradient ascent with a stochastic perturbation (random search), using constant variance,

$$\tilde{w} = w + \epsilon[\tfrac{1}{2}\nabla \log \pi(w)] + \sqrt{\epsilon} Z$$

- Or using variance inversely proportional to the target

$$\tilde{w} \quad = \quad w \quad + \quad \sqrt{\epsilon/\pi(w)} Z$$

- Approximate Invariance holds in both cases

  If $\quad w \sim \pi(w)$

  Then $\tilde{w} \sim \pi(w) e^{O(\epsilon^2)}$

Diffusions with Invariant Transitions

- Continuous-time Markov Process

$$dw_t = \mu(w_t)dt + \sigma(w_t)dB_t$$

- $\pi(w)$ is its invariant density iff μ, σ, π satisfy Kolmogorov foward equation

$$0 = -\nabla \cdot (\mu(w)\pi(w)) + \tfrac{1}{2}\nabla \cdot \nabla(\sigma^2(w)\pi(w)$$

- Discrete time transition then has π as its approximate invariant

$$\tilde{w} = w + \epsilon\mu(w) + \sqrt{\epsilon}\sigma(w)Z$$

- reference solutions for $\big(\mu(w), \sigma^2(w)\big)$

$$\left(\tfrac{1}{2}\nabla \log \pi(w), 1\right) \qquad \text{or} \qquad \left(0, \tfrac{1}{\pi(w)}\right)$$

- Evolution of the density $p(w,t)$ of w_t starting from $p(w,0) = p_0(w)$

$$\frac{\partial}{\partial t}p(w,t) = -\nabla \cdot (\mu(w)p(w,t)) + \tfrac{1}{2}\nabla \cdot \nabla(\sigma^2(w)p(w,t))$$

- Convergence to $\pi(w)$ can be slow

Recap

- Markov Chain with time-homogeneous transition can be slow to converge
- Instead: we advocate more ambitious consideration of time-inhomogeneous transitions,

  $$p_t(\tilde{w}|w)$$

  designed in discrete-time for a sequence of increasing gains γ_t to move the density from

  $$w_{t-1} \sim p_{\gamma_{t-1}}(w)$$

  to

  $$w_t \sim p_{\gamma_t}(w)$$

- A chain which achieves these aims is called an $adaptive\ annealing$

Advantages of Adaptive Annealing

- More explicit control of the sequence of densities, starting from $p_0(w)$ and tracking $p_{\gamma_t}(w)$
- Permits linear scheduling of the gain $\gamma_t = \epsilon t$, for $t = 0, 1, \ldots, \gamma_{final}/\epsilon$
- The required move bias or drift is not based solely on an uphill (gradient) direction
- Allows substantial downhill movement if it is what it takes to track the density
- General equations available for the transitions based on the continuous-time case
- Not all objective functions permit clean evaluation of these transitions
- Effort required to develop the general solution for special cases of interest

Simulated Annealing

- Kirkpatrick et al (1983)
- Aim is to track the density $p_{\gamma_t}(w)$ proportional to $e^{\gamma_t L(w)} p_0(w)$
- The reciprocal of the gain γ_t is called the temperature, in analogy with models from physics and metallurgy
- The transition rule $p_t(\tilde{w}|w)$ is Metropolis for the target p_{γ_t}
- The transition rule is invariant if the target were not changing
- But it does not take w_{t-1} with density $p_{\gamma_{t-1}}$ into w_t with density p_{γ_t}
- This shortcoming means existing theory for simulated annealing is restricted to very slowing changing gains
- Logarithmic scheduling: $\gamma_t = \frac{1}{b} \log(1+t)$, requires exponential time as a function of the dimension d to get to a suitably high final gain

Naive Approximate Diffusion

- Analogous to Simulated Annealing
- One seeks to increment the gain $\tilde{\gamma} = \gamma + \delta$
- But the drift is designed for invariance and does not account for the changing target

$$\tilde{w} = w + \epsilon[\tfrac{1}{2}\nabla \log p_{\tilde{\gamma}}(w)] + \sqrt{\epsilon}Z$$

- If $w \sim p_\gamma(s)$ then

$$\tilde{w} \sim p_\gamma(s) \exp\{-\epsilon\delta[\nabla \cdot \nabla L(w) + \gamma||\nabla L(w)||^2] + O(\epsilon^2\delta)\}$$

- The exponent is not a multiple of $(L(w) - c)$ so it fails to tract the density with increased gain

Adaptive Annealing

- Properly modify the drift to produce the desired change in the distribution

$$\tilde{w} = w + \epsilon[\tfrac{1}{2}\nabla \log p_\gamma(w) - G(w)] + \sqrt{\epsilon}Z$$

- Now if $w \sim p_\gamma(s)$ then

$$\tilde{w} \sim p_\gamma(s) \exp\left\{\epsilon[\nabla \cdot G(w) + G(w) \cdot \nabla \log p_\gamma(w)] + O(\epsilon^2)\right\}$$

- The approximation holds except for w in a negligible set
- The constant in the remainder depends on bounds on derivatives of L and derivatives and moments of G
- Choose $G(w)$ so that the expression in the exponent matches $L(w) - c$, so that the density is indeed changed in the manner desired

Choice of the Modifier $G(w)$

- Choose $G(w)$ so that the expression in the exponent matches $L(w) - m_\gamma$
- Here $m_\gamma = E_{P_\gamma} L(w)$ is the proper constant adjustment in the exponent, updating the normalizing constant.
- Consequently pick $G(w) = G_\gamma(w)$ to solve the PDE

$$[\nabla \cdot G(w) + G(w) \cdot \nabla \log p_\gamma(w)] \quad = \quad [L(w) - m_\gamma]$$

- or equivalently,

$$\nabla \cdot [G(w) p_\gamma(w)] \quad = \quad [L(w) - m_\gamma] p_\gamma(w)$$

- Recalling that $p_\gamma(w) = e^{\gamma L(w)} p_0(w)/c_\gamma$, we recognize the right side as $\partial p_\gamma(w)/\partial\gamma$, capturing the desired change in the target density

Accumulating the error

- For discrete-time Markov chain $w_0, w_1, \ldots, w_T$ with the drift modifier G_{γ_t}
- For the linear gain schedule $\gamma_t = t\epsilon$ and $T = \gamma/\epsilon$ at $\gamma = \gamma_{final}$
- We approximately track the sequence of target densities $p_{\gamma_t}(w)$ with accumulated L_1 error of order

$$O(T\epsilon^2) = O(\gamma\epsilon)$$

- Hidden constants C depend on bounds on derivatives of L and G and can be large in some cases
- The error is made small by choosing sufficiently small $\epsilon = o(1/\gamma)$
- How many steps T are required?

  To have small L_1 error for p_γ as the approximate density of w_T we need T to be large compared to γ^2

Solving for $G(w)$

- Drift modifier $G(w) = G_\gamma(w)$ required to track the distributions $p_\gamma(w)$
- Solve for $G(w)$ or equivalently $H(w) = G(w)p_\gamma(w)$ such that for a known $f(w) = [L(w) - m_\gamma]p(w)$ with zero integral, we have satisfaction of the divergence equation

$$\nabla \cdot H(w) = f(w)$$

- The traditional solution of such a PDE is obtained in the form $H(w) = \nabla h(w)$ where this h satisfies the associated Poisson equation, that the Laplacian of h equals the specified function:

$$\nabla \cdot \nabla h(w) = f(w)$$

- The solution, also characterized as the function minimizing $\int ||\nabla h(w)||^2 dw - \int h(w)f(w)dw$, is known to be obtained for by convolving $f(w)$ with the Green's function $Green(w)$ which is multiple of $1/||w||^{d-2}$ for dimensions $d > 2$. Consequently the ideal drift modifier is

$$G(w) = \frac{1}{p(w)} \int \nabla Green(w - \tilde{w})[L(w) - m]p(\tilde{w})d\tilde{w}$$

- In the $1-$dimensional case one has the simple solution in which $H(w)$ is obtained by single variable integration of $(L(w) - m)p(w)$ up to the point w.
- The resulting $G(w)$ is bounded as long as m is the mean of $L(z)$ and $p(w)$ has suitably rapid decay of its tails, as in the case that the reference $p_0(w)$ is a Gaussian.
- The challenge is to determine which problems of interest permit the modifier $G(w)$ to be computed

Diffusions with inhomogeneous transitions

- Continuous-time Markov Process with time-varying drift and variance functions

$$dw_t = \mu_t(w_t)dt + \sigma_t(w_t)dB_t$$

- Initialized by $w_0 \sim p_0$, the density function $p_t(w)$, drift $\mu_t(w)$ and variance function $\sigma_t^2(w)$ are related by the Fokker-Plank or Kolmogorov equation

$$\frac{\partial}{\partial t} p_t(w) = -\nabla \cdot (\mu_t(w) p_t(w)) + \frac{1}{2} \nabla \cdot \nabla (\sigma_t^2(w) p_t(w))$$

- For a given $p_t(w)$ we want to track for $t > 0$, solutions for $\mu_t(w)$ and $\sigma_t^2(w)$ decompose into reference solutions (for which the right side is 0) and a modifier $G_t(w)$. In particular, we may set $\sigma_t(w) = \sigma_{p_t}^{ref}(w)$ and

$$\mu_t(w) \quad = \quad \mu_{p_t}^{ref}(w) \quad - \quad G_t(w)$$

where $\mu_p^{ref}(w)$ and $\sigma_p^{ref}(w)$ are drift and variance functions for which p is invariant and where the modifier satisfies the divergence equation:

$$\nabla \cdot [G_t(w) p(w)] \quad = \quad \frac{\partial}{\partial t} p_t(w)$$

Ridge Superposition Optimization

- Given smooth univariate functions $f_1(z_1), f_2(z_2), \ldots, f_n(z_n)$ and given vectors $x_1, x_2, \ldots, x_n$, each in R^d, we want to optimize

$$L(w) = \sum_{i=1}^{n} f_i(x_i \cdot w)$$

- These f_i may be built from a sinusoidal, sigmoidal, or ridgelet function as arise in trigonometric expansions, neural nets, or multivariate wavelet analysis
- Statistical learning motivation from classification and regression problems
- Have data (x_i, y_i) for $i = 1, 2, \ldots, n$ and a fixed bounded smooth function $\psi(z)$ (such as $\sin(z)$ or $\tanh(z)$)
- We seek to optimize $\sum_{i=1}^{n}(y_i - \psi(w \cdot x_i))^2$ or to maximize

$$\sum_{i=1}^{n} y_i \psi(w \cdot x_i)$$

- Provably statistically accurate linear combinations of such ridge functions (Jones 92, Lee, Bartlett, Williamson 96, Barron, Cohen, Dahmen, Devore 07) using a greedy strategy
- Requires repeatedly performaing the following optimization task:

  Given the residual R_i from a fit $\sum_{j=1}^{k-1} \beta_j \psi(w_j \cdot x)$ using $k-1$ such terms, choose the internal weights $w = w_k$ of the kth term so as to do achieve within a value within a constant factor of the maximum of the objective function

$$L(w) = \frac{1}{n} \sum_{i=1}^{n} R_i \psi(w \cdot x_i)$$

- We recognize this optimization to be of ridge superposition form

Ridge Superposition Sampling

- Target density $p_\gamma(w)$ proportional to

$$e^{\gamma \frac{1}{n} \sum_{i=1}^n f_i(x_i \cdot w)} p_0(w)$$

- Evaluation of modifiers $G_\gamma(w)$ appears to be a mess
- For a smoothed version of the problem, variable augmentation appears to considerably clean things up
- Instead of constraining z_i to equal $x_i \cdot w$ we relax this using a narrow Gaussian to keep them close to each other.
- Then move in the n-dimensional space of the z rather than the d dimensional space of the w
- Augmented variable joint density

$$p_\gamma(w, z) = \frac{1}{c_\gamma} e^{\gamma \frac{1}{n} \sum_{i=1}^n f_i(z_i)} p_0(w) e^{-\frac{1}{2\delta^2} \sum_{i=1}^n (z_i - x_i \cdot w)^2} / (2\pi\delta^2)^{n/2}$$

Simplifying properties of augmentation

- Density for z is

$$p_\gamma(z) = \frac{1}{c_\gamma} e^{\gamma \frac{1}{n} \sum_{i=1}^n f_i(z_i)} p_0(z)$$

- It is of the form we have been studying, but now with an additive objective function The p_0 is now a Gaussian with a covariance that captures that z is near the linear space spanned by the x
- Helps decouple the Poisson equation to determine properties of the modifier $G_\gamma(z)$ and perhaps approximately solve for it
- The conditional density for z given w is of product form (conditionally independent) with a simple evolution in γ
- The conditional density for w given z is a fixed Gaussian
- The marginal density for w is of the form

$$p_\gamma(w) = \frac{1}{c_\gamma} e^{\gamma \frac{1}{n} \sum_{i=1}^n \tilde{f}_i(x_i \cdot w)} p_0(w)$$

- Here $\tilde{f}_i$ is a smoothed version of f_i, approximately the result of convolution with a narrow Gaussian of standard deviation δ.

Adaptive Gibbs Annealing

- Gibbs sampling alternates between picking z and w in a chain with the given conditionals
- For the ridge superposition model, such Gibbs sampling can be readily implemented
- Using transitions based on the rule invariant for a particular γ suffers from the same difficulties we have discussed when initializing with p_0 different from the final target
- Adaptive Gibbs Annealing is conceivable, solving for a change factor in the density that allows the density to track p_{γ_t}
- There is a similar second order PDE for the form of this change factor

Summary

- Adaptive annealing chooses a sequence of transition densities designed to approximately track a given sequence p_{γ_t} of densities designed for approximate optimization of an objective function for moderately large γ
- The solution requires a function G_t which modifies the drift
- It can be characterized by a first order PDE
- Ridge superposition objective functions permit variable augmentation which simplifies the structure of the problem
- Hopefully you will find this a useful way of better understanding Markov chains for sampling and optimization
- Opportunity for additional developments