Adaptive Annealing

Andrew R. Barron and Xi Luo

Abstract— **We provide a new approach to stochastic optimization of smooth functions and give attention in particular to the optimization of superpositions of ridge functions.**

I. INTRODUCTION

Let $L(w)$ be a bounded and smooth function of w in $\mathbb{R}^d$ that we seek to optimize. Of particular interest is the case that such functions are built by composition of simpler functions. Indeed, let univariate functions $f_1, \ldots, f_n$ be given along with vectors $x_1, \ldots, x_n$ in $\mathbb{R}^d$. Motivated by problems of function estimation for statistical learning, our interest is in optimization of superpositions of ridge functions, which take the form

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

for w in $\mathbb{R}^d$ with dimension d possibly large. We require the f_i to be bounded and to have at least two bounded derivatives. It is sufficient for our purposes to find a $\hat{w}$ producing a value $L(\hat{w})$ within a constant factor of the maximum value of $L(w)$. Moreover, to avoid an overly large magnitude w we are content with a small $\lambda > 0$ to optimize the Lagrangian

$$L(w) - \lambda\|w\|^2. \tag{2}$$

Our motivation for objective functions of the form (1) comes from statistical regression and classification. The data are x_i, y_i, for $i = 1, \ldots, n$, where for case i, the vectors x_i are observed values of d input variables and the y_i are observed response values y_i, which for the classification problem are labels in $\{-1, +1\}$. We have a parameterized family of predictors, called discriminant functions in the classification setting, which take the ridge form $\psi(x \cdot w)$, such as a sigmoid, a sinusoid, or a ridgelet. Here the w is a parameter vector of internal weights with which the inputs are linearly combined. First, for a single such predictor, to obtain an empirically driven choice of w, consider the least squares problem of optimization of $\sum_{i=1}^n (y_i - \psi(x_i \cdot w))^2$ or the closely related problem of maximization of $\sum_{i=1}^n y_i \psi(x_i \cdot w)$. We recognize both of these objective functions to be superpositions of ridge functions.

These problems are known to be computationally difficult. Even for smooth sigmoidal functions such as $\psi(z) = \tanh(z)$, it is known that $\sum_{i=1}^n (y_i - \psi(x_i \cdot w))^2$ can have exponentially (in d) many minima. Moreover, with the step sigmoid $\psi(z) = \text{signum}(z)$, except in the special case of linear separability, exact optimization is known to be *NP*-complete. Indeed, minimizing the squared error is for the signum function the same as the maximization of $\sum_{i=1}^n y_i \psi(x_i \cdot w)$. This discriminant optimization problem is also called the weighted hemisphere problem and it is known to be NP-complete as shown in [1] and discussed further in [2]. This problem is also the single neuron case of multi-neuron perception optimization. Indeed, one can more flexibly consider fitting functions which take the form of a linear combination of ridge terms. In such a setting, NP-completeness results are available both for the step sigmoid [3] and for a ramp sigmoid (a bounded piecewise linear function) provided $\|w\|$ is unconstrained [4], see also [5].

For the logistic sigmoid as arises in classification by logistic regression, a likelihood-based formulation of the objective function is concave and hence readily maximized. However, this concavity is lost when one seeks multi-term discriminant functions as arise in a mixture model.

Altogether, computation of parameter estimates for fitting ridge terms is a tricky task. Nevertheless, with a fixed smooth function ψ, approximate optimization to achieve within a constant factor of the maximum $(1/n)\sum y_i \psi(x_i \cdot w)$ (with a $\lambda\|w\|^2$ penalty) and a similar optimization for the multi-term case developed further below are the sorts of optimization problems for which we seek additional attention. In particular we explore what may be possible by stochastic optimization algorithms.

Most of what we have to say in this paper is for general smooth objective functions. We develop a class of algorithms we call adaptive annealing. In general, implementation of adaptive annealing requires solving a differential equation for what we call a modifier. We do not expect this solution to be readily available for general optimization tasks. We initiate exploration of whether solution is possible for specific forms of objective functions such as superpositions of ridge functions, perhaps with a suitable enlargement of the state space.

We follow a familiar tactic of stochastic search, seeking to sample from the distribution with density $p_\gamma(w)$ proportional to

$$e^{\gamma L(w)} p_0(w) \tag{3}$$

where $p_0(w)$ is a normal reference density. Though the development is readily adjusted to allow any normal density for p_0, for simplicity now we take it to be a standard multivariate Normal$(0, I)$. Then the mode of the density $p_\gamma(w)$ is the maximizer of $L(w) - \lambda\|w\|^2$ for $\lambda = 1/2\gamma$. An additional benefit of such sampling is that it provides for Monte Carlo evaluation of the Bayes estimate (the posterior mean of $\psi(x \cdot$

A. Barron is with Department of Statistics, Yale University, New Haven, CT 06520 andrew.barron@yale.edu

X. Luo is with Department of Statistics, Yale University, New Haven, CT 06520 xi.luo@yale.edu

w)) with $p_\gamma(w)$ as the posterior distribution. For instance, with real-valued y_i modeled as Normal$(\psi(x_i \cdot w), \sigma^2 I)$ with the normal prior p_0, we have that the posterior takes the indicated form with $L(w) = -\frac{1}{n}\sum_{i=1}^n (y_i - \psi(x_i \cdot w))^2$ and $\gamma = n/(2\sigma^2)$.

For optimization, the idea of sampling from such a density is that, if γ is large then the distribution of w is highly likely to have $L(w) - \lambda\|w\|^2$ near the maximum. An information-theoretic characterization of that property is given in Lemma 1 below, showing that γ of order not smaller than $d \log d$ is sufficiently large for our purposes. Such joint densities $p_\gamma(w)$ vary on an exponential scale, that is, the ratio of values of $p_\gamma(w)$ at two distinct w of the same norm $\|w\|$ can be of order $e^{C\gamma}$ which is exponentially large in γ and hence superexponentially large in d. Consequently, direct acceptance/rejection sampling from a reference distribution is highly unlikely to succeed.

For sampling from distributions of several variables, classical strategies consist of Markov chain methods with time-homogeneous transition rules, including the Metropolis algorithm [6], [7] and the Gibbs sampler [8], [9], and time-inhomogeneous Markov chain methods, including Kirkpatrick's simulated annealing [10] and more recently developed particle filters [11] and population Monte Carlo [12]. As we shall discuss, Markov chain strategies may also use a discrete-time approximation to stochastic diffusion [13], [14] designed to produce the desired distribution.

The general idea of all these stochastic strategies is to produce values $w_0, w_1, w_2, \ldots, w_T$ in succession by drawing from the sequence of transition distributions of the Markov chain. One initializes with a density p_0 (such as a multivariate normal) that readily permits direct sampling of w_0. This p_0 may be far from the ultimate target p_γ. To be successful in a computationally manageable time T, one needs the distribution of w_T sufficiently close to that of $p_\gamma(w)$. Likewise, for Monte-Carlo estimates one needs the perhaps weaker property that $\frac{1}{T}\sum_{t=1}^T \psi(x \cdot w_t)$ be close to the desired expectation using $p_\gamma(w)$.

The basis of the above-mentioned Markov chain methods in the time-homogeneous transition case is an invariance property: namely, if the distribution of w_{t-1} were the targeted $p_\gamma(w)$ then so would be the distribution of w_t (and the targeted distribution is the unique one for which this invariance holds). With arbitrary initial distribution $p_0(w)$, the distribution of w_t will evolve in some manner. The main conclusion known in this case is that the target distribution is the large t limit. This convergence however can be excruciatingly slow for multi-modal target densities. Existing positive results providing polynomial (in d) bounds on the number of time steps sufficient to produce suitably accurate distributions are limited to log-concave (strongly unimodal) target densities [15], [16], [17] (or to sampling from chains on discrete graphs with limited bottleneck [18], [19]) and are based on demonstration that the second largest magnitude eigenvalue is less than 1 by a not exponentially small amount.

The problem for multi-modal densities is the extreme time it can take for w to move from one mode to another if it needs to go down through a valley of exponentially lower density. As we shall recall, time-homogenous Markov Chains, designed to be invariant for a specified density, have a bias toward uphill moves which is the source of some of this difficulty in moving from one mode to another. Simply inspecting the derivative of the log-density at the current position is enough to build an invariant transition, but as we shall see, it does not give what is needed to produce a specified sequence of distributions approaching the target.

To overcome these difficulties, multiple paths of time-inhomogeneous Markov chains deserve further consideration and further development. These developments should include incorporation of drift or bias terms designed to be whatever it takes to track the changing distribution. In particular, this means a mean direction of move that is not necessarily in the gradient (uphill) direction.

The idea of simulated annealing is to evolve the distribution of w_t toward the target. Starting with a broad distribution $p_0(w)$ at time $t = 0$, the aim is for the distribution of w_t to track $p_{\gamma_t}(w)$ where γ_t is scheduled to increase (hopefully not too slowly) in t. We argue that the Markov chain associated with traditional simulated annealing falls short in these objectives. In particular, heretofore, the basic step of simulated annealing is to use a transition distribution $p_{\gamma_t}(w_t|w_{t-1})$ formulated (as in one step of the Metropolis algorithm) in a somewhat peculiar way, such that if somehow w_{t-1}, ideally distributed accords to $p_{\gamma_{t-1}}$, did on the contrary manage already to have the distribution p_{γ_t}, then so would w_t. However, as we show in Lemma 3 in our setting, such transition rules, if w_{t-1} were distributed as $p_{\gamma_{t-1}}$, actually lead to a w_t whose marginal density differs from what we want by a factor which is the exponential of a differential expression that we identify. For this differential expression to be close to what we want, one needs γ_t very close to γ_{t-1} or one needs to modify the transition distribution as we discuss below. Accordingly, for traditional similated annealling one needs a scheduling of γ_t in which it changes very slowly to have the distribution track p_{γ_t}. Indeed, [20], [14] show that a logarithmic growth of γ_t is sufficient and such slow growth is known to be necessary in some cases. To reach a targeted γ of at least d, logarithmic scheduling requires a number of time steps which is exponential in d. Thus existing simulated annealing results point toward its failure to track p_{γ_t} closely enough to allow the polynomial growth of γ_t necessary for practical use.

Nevertheless, the general idea of annealing is on track, provided suitable transition distributions can be identified. Here we introduce what we call *adaptive annealing* in which by solving the appropriate differential expression one finds a transition density (as a modification of traditional choices) such that to greater accuracy the sequence of densities for w_t track the targeted $p_{\gamma_t}(w)$.

Markov chain methods tend to be either of a local move type (as in the original Metropolis formulation or approximate diffusion) or of the Gibbs type (in which one defines transitions via auxiliary variables which augment the state space). We provide versions of adaptive annealing for both cases. For local move chains a discrete time approximation to a diffusion with time-homogeneous transitions sets $p(w_t|w_{t-1})$ to have mean $w_{t-1} + \frac{\delta^2}{2}\nabla \log p_\gamma(w_{t-1})$ and covariance $\delta^2 I$. Instead, to evolve the distribution (starting from $p_0(w_0)$), our adaptive annealing sets the transition to have mean $w_{t-1}+\delta^2[(1/2)\nabla \log p_{\gamma_t}(w_{t-1})+G_t(w_{t-1})]$ with G_t designed to produce the desired evolution of the density p_{γ_t}. This is a time-inhomogeneous Markov Chain in discrete-time for which the distribution approximates a continuous-time stochastic diffusion with corresponding time-varying drifts, see section IV. Thus a Kolmogorov forward equation (also called the Fokker-Planck equation) specifies how p_{γ_t} evolves in terms of the choice of G_t. We turn the idea around, using the specified p_{γ_t} in this equation to seek a suitable choice for G_t. This is a simple and interesting idea, though direct evaluation of G from the differential equation is elusive for the types of optimization of interest to us.

Adaptive annealing of Gibbs-type chains is explored preliminarily in section VI. For Gibbs chains the idea is that one introduces an auxiliary variable z and joint density $p_\gamma(w, z)$ for which the desired $p_\gamma(w)$ is its marginal and for which the conditionals $p_\gamma(w|z)$ and $p_\gamma(z|w)$ take special forms that allow for more direct sampling. When such conditionals are available, repeated alternate sampling from these pairs of conditionals provides a Markov chain for which $p_\gamma(w, z)$ is the invariant target. Once again, unmodified convergence can be slow when the distribution of w_0 or z_0 is initialized, naturally, with other than the target p_γ. Recognizing the discrepancy between the initial distribution and the target we again aim to take control of the path of distributions to take us to that target.

Thus we advocate Gibbs annealing in which we track the sequence of densities $p_{\gamma_t}(w, z)$ for $t = 0, 1, 2, \ldots, T$. Again, if at best w_{t-1} is distributed with density $p_{\gamma_{t-1}}(w)$, then unfortunately the transition $p_{\gamma_t}(z|w)$ of a naive Gibbs annealing does not make the marginal for the resulting z be what we want. Accordingly, we introduce adaptive Gibbs annealing in which the transition density is designed to achieve our aims. In particular we introduce a factor by which to modify the transition $p_{\gamma_t}(z|w)$ in acceptance/rejection sampling. This leads once again to a new density for z_t, w_t that differs from what we want by a factor that is the exponential of a differential expression. The aim then is to solve this differential expression to provide such a modifying function that will produce indeed the desired $p_{\gamma_t}(z_t, w_t)$ to greater precision.

After developing these ideas in the indicated sections we explore in particular the ridge function case. This provides a natural setting for introduction of the auxiliary variable z. Finally, we point toward possibilities to solve for the transitions that would provide practical stochastic optimization for the problems of interest.

II. IMPLICATIONS FOR FUNCTION FITTING

The classical statistical function estimation problem is to seek a function $\hat{f}(x)$ to estimate the unknown underlying function $f(x)$ of which we have (possibly noisy) observations $y_1, \ldots, y_n$ at corresponding inputs $x_1, \ldots, x_n$ in $\mathbb{R}^d$.

A widely studied class of estimators $\hat{f}(x)$ takes the form of a composition of finitely many terms chosen from a dictionary of candidate terms. Whereas in low-dimensional (small d) settings, one may arrange for a manageable size dictionary (e.g. the union of various wavelet bases of interest); in high-dimensional cases one tends to have an intrinsically exponential number of candidate terms (dictionaries with minimal δ-covering numbers of order $(1/\delta)^d$). The dictionary of candidate terms may take a parametric form $\{\psi(x \cdot w) : w \in \mathbb{R}^d\}$ with a continuous parameter w. For instance this includes sinusoidal and sigmoidal models [21], [22] and ridgelets [23].

Function fits take the form

$$\hat{f}_m(x) = \sum_{j=1}^{m} \beta_j \psi(x \cdot w_j). \tag{4}$$

The computationally difficult full least squares solution seeks $\beta_1, \ldots, \beta_m$ (each in $\mathbb{R}$) and $w_1, \ldots, w_m$ (each in $\mathbb{R}^d$) to minimize

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{m} \beta_j \psi(x_i \cdot w_j))^2.$$

Fortunately a well-studied computational shortcut [24], [22], [25], [26] shows that statistically accurate fits can be obtained iteratively by setting

$$\hat{f}_k(x) = (1 - \alpha)\hat{f}_{k-1}(x) + \beta\psi_k(x \cdot w), \tag{5}$$

with $\hat{f}_0(x) = 0$. Here α and β are fit by least squares and the internal parameter vector $w = w_k$ is chosen to make

$$\frac{1}{n}\sum_{i=1}^{n} R_i \psi(x_i \cdot w) \geq \frac{1}{C} \max_w \{\frac{1}{n}\sum_{i=1}^{n} R_i \psi(x_i \cdot w)\}. \tag{6}$$

for given $C \geq 1$, where $R_i = R_{i,k-1} = y_i - \hat{f}_{k-1}(x)$ and where penalties on $\|w\|^2$ may be incorporated. We recognize this optimization to be of the desired form with $L(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(x_i \cdot w)$ with $f_i(z) = R_i \psi(z)$.

With a suitable definition of the variation $V(f)$ of f with respect to the dictionary, the resulting $\hat{f}_m(x)$ is proved in [26] to satisfy

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_m(x_i))^2 \leq \min_f \{\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \frac{4C^2V^2(f)}{m}\}. \tag{7}$$

with similar conclusion given for the risk quantifying the accuracy with which $\hat{f}$ generalizes.

Thus development of an algorithm to approximately optimize objective functions of ridge superposition form as expressed in (6) will provide a means to produce accurate fits of the ridge superposition form (4). Note the duality. The functions we fit are ridge superpositions in x while the objective function is a ridge superposition in w. The greedy shortcut expressed in (5) is what reduces what would be a much more complicated optimization with multiple w's into the simpler objective function with a single w.

III. HOW BIG SHOULD BE THE GAIN γ?

We show for objective functions with bounded gradient that for w drawn from p_γ, the expected value of $L(w)$ is nearly maximal. In particular, compared to the value $L(w^*)$ achieved by any particular w^*, the expectation

$$L_{mean} = \int L(w)p_\gamma(w)dw \tag{8}$$

is not less than $L(w^*)$ by more than $\frac{d}{\gamma}\log\frac{A\gamma}{d}$, where A depends on $\|w^*\|$ and the maximal gradient of L. Consequently, γ slightly larger than d by a log factor is sufficient to obtain usefully accurate samples for approximate maximization.

Let $L(w)$ be our objective function for $w \in \mathbb{R}^d$. Let $\nabla L(w)$ be its gradient, let $\|\nabla L(w)\|$ be its Euclidean norm, and let $\|\nabla L\|_{max} = \sup_w \|\nabla L\|$ be the maximum gradient norm. We have $p_\gamma(w) = e^{\gamma L(w)}p_0(w)/c_\gamma$ where $c_\gamma = \int e^{\gamma L(w)}p_0(w)dw$ and we choose $p_0(w) = e^{-\|w\|^2/2}/(2\pi)^{d/2}$.

Lemma 1: $L(w)$ has nearly maximal expectation: for any w^* and any $\gamma > 0$,

$$\begin{aligned} L_{mean} \geq L(w^*) - \Bigg\{ &\frac{d}{\gamma}\log(\frac{\gamma a\|\nabla L\|_{max}}{d}) \\ &+ \frac{1}{2\gamma}(\|w^*\| + \frac{d}{\gamma\|\nabla L\|_{max}})^2 \Bigg\} \end{aligned} \tag{9}$$

where $a = eV_d^{1/d}/\sqrt{2\pi}$ with V_d the volume of the unit ball in $\mathbb{R}^d$.

Remark 1: Consequently

$$L_{max} \geq L(w^*) - \tfrac{d}{\gamma}\log\tfrac{\gamma A}{d}, \tag{10}$$

with A near $a\|\nabla L\|_{max}e^{\frac{1}{2d}\|w^*\|^2}$. Hence we find $\gamma \geq \frac{2ed}{L(w^*)}\max\{e, \log\frac{2Ad}{L(w^*)}\}$ is sufficient to have

$$L_{mean} \geq \tfrac{1}{2}L(w^*). \tag{11}$$

So we see that, on the average for w drawn from p_γ, the value of $L(w)$ can be arranged to be within a constant factor of the maximum. The reference w^* may be taken to be have maximum $L(w^*) - \lambda||w^*||^2$. Moreover, using a Markov inequalities we can see that $L(w)$ is within a constant factor of the maximum with high probability.

Proof: First we find that

$$L_{mean} \geq \tfrac{1}{\gamma}\log c_\gamma. \tag{12}$$

This can be seen either by noting that $\log c_\gamma$ has a derivative $L_\gamma = \int L(w)p_\gamma(w)dw$ which is increasing in γ, so $\frac{1}{\gamma}\log c_\gamma$ which is an average from 0 to γ of that derivative, is less than the value of the derivative at the upper endpoint γ, which is $L_\gamma = L_{mean}$. Alternatively, one finds that the Kullback divergence between p_γ and p_0 is equal to the difference $\gamma L_\gamma - \log c_\gamma$ which also shows (12) by the non-negativity of divergence. Next, adding and subtracting $L_* = L(w^*)$ in the exponent of the integral defining c_γ, we then have

$$L_\gamma \geq L_* + \frac{1}{\gamma}\log\int e^{\gamma(L(w)-L_*)}p_0(w)dw. \tag{13}$$

We lower bound this further by restricting the integral to the ball B centered at w^* with radius ϵ and by using

$$|L(w) - L(w^*)| \leq \|w - w^*\|\|\nabla L\|_{max}. \tag{14}$$

Hence

$$L_\gamma \geq L_* - \epsilon\|\nabla L\|_{max} + \frac{1}{\gamma}\log\int_B p_0(w)dw. \tag{15}$$

Moreover

$$\int_B p_0(w)dw \geq (\frac{1}{\sqrt{2\pi}})^d e^{-\frac{1}{2}(\|w_*\|+\epsilon)^2}\epsilon^d V_d. \tag{16}$$

Thus

$$\begin{aligned} L_\gamma \geq L_* - \Bigg\{ &\epsilon\|\nabla L\|_{max} - \frac{d}{\gamma}\log\epsilon + \frac{1}{2}(\|w_*\| + \epsilon)^2 \\ &+ \frac{1}{\gamma}\log(\sqrt{2\pi})^d/V_d \Bigg\}. \end{aligned} \tag{17}$$

Picking $\epsilon = d/(\gamma\|\nabla L\|_{max})$ completes the proof. ■

IV. APPROXIMATE DIFFUSION

In this section we set up the principles of approximate diffusion including the choice of the ideal drift function to track a specified sequence of distribution. The terminology of simulated annealing for optimization arises in analogy with the physical annealing process of metallurgy. Likewise, stochastic diffusion models have roots in physical phenomena of particle diffusion. The behavior of these processes have been modeled by physicists and mathematicians using partial differential equations that describe time evolutions of the distributions. A general diffusion w_t, $0 \leq t \leq T$, with continuous paths is denoted by the stochastic differential equation

$$dw_t = \mu(w_t, t)dt + \sigma(w_t, t)dB_t \tag{18}$$

where B_t is the standard Brownian motion process. Here the vector-valued function $\mu(w,t)$ is a possibly time-varying local drift function and we take $\sigma^2(w,t)$ to be a non-negative scalar-valued local variance function.

Initialized by a distribution with density $p(w,0) = p_0(w)$, such processes w_t have density functions $p(w,t)$ for $t \geq 0$. The time evolution of the probability density function is governed by a PDE named the Fokker-Planck equation after its first developers [27], also called a Kolmogorov forward equation, see for instance [14]:

$$\frac{\partial p(w,t)}{\partial t} = -\nabla\cdot(\mu(w,t)p(w,t)) + \frac{1}{2}\nabla\cdot\nabla(\sigma^2(w,t)p(w,t)). \tag{19}$$

Here ∇ denotes the gradient operator (the vector of derivatives with respect to w). Its dot product with a vector-valued function is the divergence and its dot product with itself is the Laplacian operator. Let $(\mu(w), \sigma^2(w)) = (\mu_{p,ref}(w), \sigma^2_{p,ref}(w))$ be a reference solution for which a density p is invariant. That is

$$0 = -\nabla\cdot(\mu(w)p(w)) + \frac{1}{2}\nabla\cdot\nabla(\sigma^2(w)p(w)). \tag{20}$$

A traditional choice is $(\mu(w), \sigma^2(w)) = (\frac{1}{2}\nabla\log p(w), 1)$ or any non-negative constant multiple thereof. This is the rule that maintains a drift $\frac{1}{2}\nabla \log p(w)$ in the gradient direction, pointing toward higher value of $p(w)$.

Less frequently used, but also possible is $(\mu(w), \sigma^2(w)) = (0, 1/p(w))$. It makes broader moves where $p(w)$ is low and narrower moves where $p(w)$ is high, in a manner such that the transition rule is invariant for p.

These solutions together with (19) may be used in two ways, in both cases assuming that we start with a density $p_0(w)$. One is to fix such $(\mu(w), \sigma^2(w))$ as a specification of a time-homogeneous transition (as in standard Markov chain samplers) and use the PDE

$$\frac{\partial p(w,t)}{\partial t} = -\nabla \cdot (\mu(w)p(w,t)) + \frac{1}{2}\nabla \cdot \nabla(\sigma^2(w)p(w,t)). \tag{21}$$

to study how the density $p(w,t)$ of w_t evolves (in particular how rapidly does it approach the specified $p(w) = p_\gamma(w)$). [Associated large deviation properties of the function $L(w)$ defining the target p_γ in the setting of these time-homogeneous chains were presented by us at a joint Statistics and Electrical Engineering seminar at the University of Illinois on February 27, 2007.]

The other tactic, as we advocate here, is to treat the density $p(w,t) = p_{\gamma_t}(w)$ as given and seek suitable functions $\mu(w,t)$ and $\sigma^2(w,t)$ for which (19) holds (rather than the other way around). This task simplifies as follows. Write the general drift function $\mu(w,t)$ in the form

$$\mu(w,t) = \mu_{\gamma_t,ref}(w) - G(w,t) \tag{22}$$

and set the variance function to be $\sigma^2(w,t) = \sigma^2_{\gamma_t,ref}(w)$ (commonly chosen to be constant as we have said). We think of $G(w,t)$ as specifying a modification of the naive choice of drift function $\mu_{\gamma_t,ref}(w)$. Now (19) simplifies to show that the modifier $G(w,t)$ leads us to achieve the target distribution $p(w,t)$, here taken to be $p_{\gamma_t}(w)$, if and only if

$$\frac{\partial p(w,t)}{\partial t} = -\nabla \cdot (G(w,t)p(w,t)). \tag{23}$$

So this provides the equation to be satisfied by the appropriate changes $G(w,t)$ to the drift function required to track the desired $p_{\gamma_t}(w)$. We see thereby that the ideal drift function for an improved stochastic search relies not only the direction of higher values of $p(w)$ (uphill), but also, via $G(w,t)$ one has means to bias movement downhill when necessary to evolve the distribution as desired.

Recall for our maximization purposes, eventually we want to the process to provide samples from a density

$$p_\gamma(w) = \frac{1}{c_\gamma} e^{\gamma L(w)} p_0(w), \tag{24}$$

and toward that end we propose to sample from a sequence of densities with increasing $0 \le \gamma_t \le \gamma$.

A discrete-time Markov chain motivated by the above consideration sets $w_0 \sim p_0(w)$ and

$$w_t = w_{t-1} + (\frac{\delta^2}{2}\nabla \log p_{\gamma_{t-1}}(w_{t-1}) - \delta^2 G_t(w_{t-1})) + \delta Z_t \tag{25}$$

where the Z_t are independent standard normals in $\mathbb{R}^d$ and $G_t(w)$ is a vector-valued function. Here $\gamma_t = t\delta^2$ for $t = 1, \ldots, T$ with $T = \gamma/\delta^2$. If $G_t(w) = G(w,t)$ is chosen to (approximately) solve the partial differential equation (23) with the targeted $p(w,t) = p_{\gamma_t}(w)$, we say that this Markov chain is an (approximate diffusion-based) *adaptive annealing*. For densities of the form (24), the required equation for $G_t(w)$ reduces to

$$\nabla \cdot (G_t(w)p_{\gamma_{t-1}}(w)) = (L(w) - L_{\gamma_{t-1}})p_{\gamma_{t-1}}(w) \tag{26}$$

or equivalently

$$\nabla \cdot G_t(w) + G_t(w) \cdot \nabla \log p_{\gamma_{t-1}}(w) = L(w) - L_{\gamma_{t-1}} \tag{27}$$

where $L_\gamma = \mathbb{E}_{p_\gamma} L(w)$, arising from $\frac{\partial}{\partial \gamma} \log c_\gamma$, and where $\nabla \log p_\gamma(w) = \gamma \nabla L(w) - w$.

The idea here is that if δ is small the distribution of w_t in the discrete-time Markov chain should be close to the target $p_{\gamma_t}(w)$. We use the total variation distance between distributions which is the L_1 distance between the density functions.

Proposition 2: The density $p_{\gamma_t,\delta}(w)$ of w_t approximately tracks $p_{\gamma_t}(w)$ in the sense that

$$\|p_{\gamma_t,\delta} - p_{\gamma_t}\|_1 = \mathcal{O}(B^2 d^4 t \delta^4) \tag{28}$$

with $\gamma_t = t\delta^2$ and hence at $T = \gamma/\delta^2$ we have

$$\|p_{\gamma_T} - p_\gamma\|_1 = \mathcal{O}(B^2 d^4 \gamma \delta^2) \tag{29}$$

where B will be specified momentarily. Thus with δ small compared to $1/(\gamma d^2 B)$ a number of steps T of order $\gamma^2 d^4 B^2$ is sufficient to have $p_{\gamma_T,\delta}$ close to p_γ.

One sufficient set of conditions is that L is bounded and has bounded derivatives of all orders, satisfying for some positive B the requirement that for all indices i_1 through i_k in $\{1, \ldots, d\}$

$$|\frac{\partial^k}{\partial w_{i_1} \partial w_{i_2} \cdots \partial w_{i_k}} L(w)| \le k! B^k. \tag{30}$$

Moreover, uniformly in t, the absolute moments of the derivatives of each of the coordinates of $G_t(w)$ is also bound by $k!B^k$ where k is the sum of the order of the derivatives and the order of the moment.

One may either require these conditions to hold for all k (in conjunction with infinite Taylor series expansion), or with different constants in the bound one may assume only that they hold for derivatives up to order 4 (though in that case one needs local domination of the derivatives to have the moment controls). We save details of such matters to later writings.

Remark 2: Similarly, the Fokker-Planck equation may be derived using Taylor expansions on a small step evolution and taking a limit as step size δ goes to 0 as in [27]. Our proof shares some general features with that development, but is customized here for our aims.

The heart of the proof of the proposition is repeated application of the following key lemma.

Lemma 3: Let $p_t(w_t|w_{t-1})$ be the transition density associated with the stochastic move defined in (25) with change function $G_t(w)$ and let

$$\tilde{p}_{\gamma_t,\delta}(w_t) = \int p_{\gamma_{t-1}}(w_{t-1}) p_t(w_t|w_{t-1}) dw_{t-1} \tag{31}$$

be the marginal density function of w_t that would arise if w_{t-1} were distributed according to $p_{\gamma_{t-1}}$. Then except for w in a set of negligible probability

$$\begin{aligned}\tilde{p}_{\gamma_t,\delta}(w) = p_{\gamma_{t-1}}(w) e^{\delta^2[G_t(w)\cdot\nabla\log p_{\gamma_{t-1}}(w)+\nabla\cdot G_t(w)]} \\ \cdot e^{\mathcal{O}[\delta^4 d^4 B^2]}.\end{aligned} \tag{32}$$

Alternatively, we may write

$$\begin{aligned}\tilde{p}_{\gamma_t,\delta}(w) = p_{\gamma_{t-1}}(w) + \delta^2 \nabla\cdot(G_t(w) p_{\gamma_{t-1}}(w)) \\ + \mathcal{O}[\delta^4 d^4 B^2].\end{aligned} \tag{33}$$

Remark 3: According to (32) one sees that even in the ideal case that the density of w_{t-1} matched the desired $p_{\gamma_{t-1}}(w)$, the density after a transition (25) does not match the desired $p_{\gamma_t}(w)$ to suitable order of accuracy, unless G_t is such that $G_t(w)\cdot\nabla\log p_{\gamma_{t-1}}(w) + \nabla\cdot G_t(w)$ is proportional to $L(w)$, that is, unless the differential equation (23) is (approximately) satisfied. As remarked in the introduction, this is the basis of our concerns leading to the advocacy of adaptive annealing.

Full proof of the Lemma and proposition are outside the scope of this conference paper. Nevertheless the following is suggestive of how it is approached.

Proof: [sketch] The density function at the next step is

$$\tilde{p}_{\gamma_t}(w_t) = \int p_{\gamma_{t-1}}(w_{t-1}) p(w_t|w_{t-1}) dw_{t-1}, \tag{34}$$

and the transition kernel $p(w_t|w_{t-1})$ is normal with mean $w_{t-1} + (\frac{\delta^2}{2}\nabla\log p_{\gamma_{t-1}}(w) + \delta^2 G(w_{t-1}))$ and variance $\delta^2 I$. Denote $\xi = w_{t-1} - w_t$. We factor out $p_{\gamma_{t-1}}(w_t)$ and rewrite the whole integral in terms of ξ, to approximate $\tilde{p}_{\gamma_t}(w_t)$ as

$$p_{\gamma_{t-1}}(w_t) \int e^{\log p_{\gamma_{t-1}}(w_t+\xi) - \log p_{\gamma_{t-1}}(w_t)} \phi(\xi|w_t) d\xi \tag{35}$$

where, using the form of the normal transition density, $\phi(\xi|w_t)$ to be a normal density that approximates $p(w_t|w_t+\xi)$. The above integral can now be viewed as an expectation of $e^{\log p_{\gamma_{t-1}}(w_t+\xi) - \log p_{\gamma_{t-1}}(w_t)}$ respect to the density function $p(\xi|w_t)$. We approximate the logarithm of this expectation by cumulants and collect dominating terms respect to δ, then we identify that the leading exponent of the factor multiplying $p_{\gamma_{t-1}}(w_t)$ is $G_t(w)^T\nabla\log p_{\gamma_{t-1}}(w) + \nabla\cdot G_t(w)$. Likewise, the exponent needed to boost $p_{\gamma_{t-1}}(w)$ to $p_{\gamma_t}(w)$ is $\delta^2(L(w) - L_{\gamma_{t-1}}(w))$. Hence we advocate $G_t(w)$ for which these two match as in (26). ■

Brief consideration of a naive approximate diffusion is warranted. By analogy with simulated annealing we could think to use for the drift function in equation (25) the choice $\frac{\delta^2}{2}\nabla\log p_{\gamma_t}(w_{t-1})$ which would be invariant at the next gain γ_t. This corresponds to a modifier G_t which is a small multiple of $-\nabla L(w)$. Such a choice leads to the differential expression $-\{\gamma||\nabla L(w)||^2 + \nabla\cdot\nabla L(w)\}$ in the exponent of the factor by which the density is modified. It appears that with this naive choice the differential expression reduces to the desired form of a multiple of $L(w)$ minus a constant only for certain quadratic objective functions. This reinforces our sense that, except for a trivial setting, use of a modifier other than the gradient of the log density is essential to track the sequence of densities p_{γ_t} when γ_t is increasing adequately fast.

Next we provide some initial discussion on how to solve for the required $G_t(w)$ from (26). For a given $p(w) = p_{\gamma_{t-1}}(w)$ we seek $G(w)$ such that

$$\nabla\cdot(G(w)p(w)) = (L(w) - L_{mean})p(w) \tag{36}$$

where $L_{mean} = \int L(w)p(w)dw$. Consider first the scalar ($d = 1$) case for which we may set

$$G(w) = \frac{1}{p(w)} \int_{-\infty}^{w} (L(\tilde{w}) - L_{mean}) p(\tilde{w}) d\tilde{w}. \tag{37}$$

Note that the integral, if extended to the whole line from $-\infty$ to $+\infty$ becomes 0 by the definition of L_{mean}, so equivalently we have

$$G(w) = -\frac{1}{p(w)} \int_{w}^{+\infty} (L(\tilde{w}) - L_{mean}) p(\tilde{w}) d\tilde{w}. \tag{38}$$

Note that with L bounded and $p(w)$ proportional to $e^{\gamma L(w) - w^2/2}$, which has Gaussian tails, the integral is seen to be controlled for large w by the tail integral of the Gaussian, which is bounded by a constant times $\frac{1}{w}e^{-w^2/2}$, in this one-dimensional case. Consequently, despite the division by $p(w)$ this $G(w)$ is seen to taper to 0 at a polynomial rate as w heads to $+\infty$ or $-\infty$.

For $d > 1$, defining the coordinates of $G(w)p(w)$ by integration of $\frac{1}{d}(L(w) - L_{mean})p(w)$ with respect to the corresponding coordinates of w would satisfy (36) but has the problem that the integral from $-\infty$ to $+\infty$ with respect to a coordinate does not give 0 (since L_{mean} is the overall mean, not arising from an integration of $L(w)p(w)$ with respect to just one coordinate) and consequently such solution leads to divergent $G(w)$ as $w \to \infty$.

Consideration of second order PDE theory suggests another avenue. One seeks $H(w) = G(w)p(w)$ as the gradient of a function $h(w)$ which solves the Poisson equation

$$\nabla\cdot\nabla h(w) = (L(w) - L_{mean})p(w),$$

and for which $h(w)$ tapers to 0 as $w \to \infty$. A solution for $h(w)$ is known to be the convolution of the right side $(L(w) - L_{mean})p(w)$ by the Green's function $Green(w)$ associated with the Laplacian on $\mathbb{R}^d$ (it is proportional to $1/\|w\|^{d-2}$ for $d > 2$) and then

$$G(w) = \frac{1}{p(w)} \int_{\mathbb{R}^d} (\nabla Green(w-\tilde{w}))(L(\tilde{w}) - L_{mean}) p(\tilde{w}) d\tilde{w}. \tag{39}$$

Once again one can show that this $G(w)$ tapers to 0 as $\|w\| \to \infty$. However, it has the problem now that it is unclear how to bound its behavior for moderate values of w due to the exponential swings in the height of $p(w) = e^{\gamma L(w)}$ with γ at least d.

Currently our preferred tactic is to seek formulation of the problem that makes repeated use of one dimensional formulations. In the case that $L(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(x_i \cdot w)$ this means working with the variable $x_i \cdot w$. This motivates our considerations in the next section.

V. Ridge Superposition Sampling

When the objective function is a ridge superposition $L(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(x_i \cdot w)$, the target density $p_\gamma(w)$ is proportional to

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

Direct evaluation of the modifiers $G_\gamma(w)$ appears to be somewhat of a mess. Nevertheless, for a smoothed version of the problem, variable augmentation appears to considerably clean things up. The idea is this: for the functions $f_i(z_i)$ 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 we have the Markov chain move in the n-dimensional space of the z rather than the d dimensional space of the w. Thus we propose the following joint density function for the augmented variables

$$p_\gamma(w, z) = \frac{1}{c_\gamma} e^{\gamma \frac{1}{n}\sum_{i=1}^{n} f_i(z_i)} p_0(w) \frac{e^{-\frac{1}{2\delta^2}\sum_{i=1}^{n}(z_i - x_i \cdot w)^2}}{(2\pi\delta^2)^{n/2}}. \tag{40}$$

Integrating out w we see that the density for z is

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

where the p_0 is now a Gaussian with a covariance that captures that z is near the linear space spanned by the variables that comprise the x.

We recognize this density to be of the form we have been studying. It is the exponential of an objective function with a Gaussian reference initial density. But now there is the considerable simplification that the objective function simply takes an additive form. This encourages determination of whether we can at least approximately solve for the associated modifiers $G(z)$. The idea being that the product structure of much of this density may help decouple the Poisson equation. This question of solving for $G(z)$ as well as other related questions for Gibbs samplers that we discuss below remain under continuing consideration.

Before preceding to a discussion of adaptive annealing of Gibbs samplers we point out some additional properties of the joint density we have here specified. Note that the conditional density for z given w is of product form $p_\gamma(z|w) = \prod_{i=1}^{n} p_{\gamma,i}(z_i|w)$, with

$$p_{\gamma,i}(z_i|w) = e^{\frac{\gamma}{n}[f_i(z_i) - \tilde{f}_i(x_i \cdot w)]} \phi_\delta(z_i - x_i \cdot w) \tag{41}$$

where $\phi_\delta(v) = \frac{1}{\sqrt{2\pi}\delta} e^{-v^2/2\delta^2}$ is the Normal$(0, \delta^2)$ density and $\tilde{f}_i(u)$ is defined via the conditional normalization $e^{\frac{\gamma}{n}\tilde{f}_i(u)} = \int_{-\infty}^{+\infty} e^{\frac{\gamma}{n} f_i(z)} \phi_\delta(z - u) dz$. This $\tilde{f}_i$ depends on γ/n and δ and it is close for small γ/n to $\int f_i(z)\phi_\delta(z - u)dz$, the convolution smoothing of f_i with the normal of standard deviation δ. Consequently, in the joint density, when we integrate out the z_i we obtain the following marginal density for w

$$p_\gamma(w) = \frac{1}{\tilde{c}_\gamma} e^{\frac{\gamma}{n}\sum_{i=1}^{n} \tilde{f}_i(x_i \cdot w)} p_0(w). \tag{42}$$

This density for w is of the form we desire, especially if δ is small. If we aim to have smooth components $\tilde{f}_i$ expressible as a convolution with a normal then one may choose f_i as a precursor that leads to such $\tilde{f}_i$. Furthermore we note that the conditional density for w given z is a fixed Gaussian. Thus Markov Chain sampling of the z followed by drawing w given z, produces an outcome w whose density is of the form we desire.

VI. Adaptive Gibbs Annealing

Desiring to sample from a target distribution $p_\gamma(w)$ one may follow the variable augmentation idea of [9]. One specifies a vector of hidden variables z which leads to a augmented state (w, z) and one specifies a joint density $p_\gamma(w, z)$ for which the desired $p_\gamma(w)$ is the marginal. Moreover for Gibbs sampling one arranges that the conditional densities $p_\gamma(w|z)$ and $p_\gamma(z|w)$ are convenient for alternating sampling. Starting from some initial z_0, w_0, and then drawing $z_1, w_2, z_2, \ldots, z_T, w_T$ in succession from the indicated conditionals is a version of Gibbs sampling [8], [9]. Once again it defines a Markov chain with time-homogeneous transitions, for which $p_\gamma(w, z)$ is invariant, but the exact sequence of distributions for w_t that results (starting from $w_0 \sim p_0$) and its rate of approach to the target $p_\gamma(w)$ is not clear.

Thus we initiate investigation of the behavior of a time-inhomogeneous version of Gibbs sampling. As before merely using $p_{\gamma_t}(z_t|w_{t-1})$ and $p_{\gamma_t}(w_t|z_t)$ as the transition rules at time t will not actually track the desired sequence of distributions p_{γ_t} with increasing γ_t. Instead we incorporate a modifier $g_t(z, w)$ and examine the transition densities proportional to $p_{\gamma_t}(z_t|w_t)e^{\epsilon g_t(z,w)}$. Thus the transition density we consider takes the form

$$\tilde{p}_{\gamma_t}(z_t|w_{t-1}) = p_{\gamma_t}(z_t|w_{t-1}) e^{\epsilon[g_t(z_t, w_{t-1}) - \tilde{g}_t(w_{t-1})]} \tag{43}$$

where $e^{\epsilon \tilde{g}_t(w)} = \int p_{\gamma_t}(z|w) e^{\epsilon g_t(z,w)} dz$ is the conditional normalizing factor. When this normalizing factor is finite for some positive ϵ, then when ϵ is small, $\tilde{g}_t(w)$ is approximately the conditional mean $\int p_{\gamma_t}(z|w) g_t(z, w) dz$ of the modifier $g_t(z, w)$.

This modified transition rule arises in multiple ways. Most directly, if $e^{\epsilon[g_t(z,w) - \tilde{g}_t(w)]}$ is as a bounded function of z, with bound B_w, say, then given $w_{t-1} = w$, one can draw z_t according to $\tilde{p}_{\gamma_t}(z_t|w)$ by repeated acceptance/rejection sampling from $p_{\gamma_t}(z_t|w)$ (in which we accept z_t when $e^{g_t(z_{t-1}, w) - \tilde{g}_t(w)}$ exceeds an independent uniform $[0, B_w]$ random variable). Note that with ϵ small, B_w is near 1 and the acceptance probability is high.

Alternatively, from a candidate $z_t \sim p_{\gamma_t}(z_t|w_{t-1})$, one may obtain a new point from the desired distribution (approximately) using the transition $z_{t,new} = z_t - \epsilon G_t(z_t|w_{t-1})$ or as in approximate diffusion, one may take $z_t - \frac{\epsilon}{2}\nabla \log p_{\gamma_t}(z_t|w_{t-1}) - \epsilon G_t(z_t|w_{t-1}) + \sqrt{\epsilon} Normal_\epsilon$, where in either of these cases $g_t(z, w) = \nabla \cdot G_t(z|w) +$

$G_t(z|w) \cdot \nabla \log p_{\gamma_t}(z|w)$. Availability the solution $G_t(z|w)$ depends on the form of $p_{\gamma_t}(z|w)$. The more direct acceptance/rejection method is preferred when solution of that differential equation for G_t is not readily accessible.

The key step of analysis of the modified transition is the following. We are given a joint distribution $p(w, z)$ with associated marginals $p(w)$ and $p(z)$ and conditionals $p(z|w)$ and $p(w|z)$. With $w \sim p(w)$ a transition using $p(z|w)$ would yield the marginal $p(z)$. We ask what happens if we instead use a modified transition.

Lemma 4: With $w \sim p(w)$ followed by z given w drawn by the modified transition $\tilde{p}(z|w) = p(z|w)e^{\epsilon[g(z,w)-\tilde{g}(w)]}$, the resulting marginal $\tilde{p}(z) = \int p(w)\tilde{p}(z|w)$ satisfies the following small ϵ approximation

$$\tilde{p}(z) = p(z)e^{\epsilon g_{mod}(z)+O(\epsilon^2)} \tag{44}$$

where $g_{mod}(z) = \int (g(z,w) - \tilde{g}(w))p(w|z)dw$ is the conditional mean of the modifier with respect to $p(w|z)$. This holds when $e^{\epsilon g(z,w)}$ has finite expectation with respect to $p(w, z)$ for some positive ϵ.

The general task then is to identify a modifier $g(z, w)$ for $p_\gamma(z|w)$ such that the expression $\epsilon g_{mod}(z)$ in the exponent approximately matches the desired change in the exponent of $p_\gamma(z)$ when we increase the gain. Applying such a modifier for a suitable increasing γ_t to the transitions from w_{t-1} to z_t, the sequence of alternating w_t and z_t, for $t = 0, \ldots, T$ is said to be an *adaptive Gibbs annealing* with distribution (approximately) tracking p_{γ_t}.

We now return our attention to the case of ridge superposition sampling. The variable augmentation we defined in the preceding section may facilitate simplification of the task of finding suitable modifiers for adaptive Gibbs annealing. Now the density of z given w depends on w only through the variables $u_i = x_i \cdot w$, which may be collected together as $u = Xw$ where X is the design matrix with rows given by the x_i. We may chose in this case for the modifier to depend on w through $u = Xw$ and write it as $g(z, Xw)$ where $g(z, u)$ is a function on $R^n \times R^n$ that needs to be specified. A special case that might be rich enough for our purposes is to have $g(z, u) = g(z)$ depend only on z. In either case the aim to to choose g so as achieve the desired $g_{mod}(z)$, which is a multiple of $f(z)$ minus a constant, where $f(z) = \frac{1}{n}\sum_{i=1}^n f_i(z_i)$.

Concerning the mechanics of performing adaptive Gibbs annealing for ridge superposition sampling, we recall that for joint density given in expression (40), the conditionals for z given w make the coordinates z_i independent with density as given in expression (41). Sampling from each of these univariate conditional densities for z_i can be performed by acceptance/rejection sampling using the Gaussian reference with mean $x_i \cdot w$ and variance δ^2. With γ/n small, we see that the factor modifying the Gaussian is close to 1, so that we will have suitably high acceptance probability for this scheme. In this manner with $w = w_{t-1}$ we may draw (repeatedly if need be) from $p_{\gamma_t}(z|w)$ and hence from the modified transition $p_{\gamma_t}(z|w)e^{\epsilon[g(z,Xw)-\tilde{g}(Xw)]}$ to obtain the vector $z_t = z$ now using acceptance /rejection of the vector as discussed earlier, where now the normalizer $\tilde{g}$ depends only of Xw. Following this draw of z_t, we can easily obtain the suitable w_t by a draw from the Gaussian conditional for w given z, or if we prefer we may bypass w_t and directly draw $u_t = Xw_t$ given z.

This Gaussian conditional for $u = Xw$ given z is the familiar posterior that arises in Bayesian regression if the hidden z_i were regarded as observed responses. This distribution is also of interest because approximate evaluation of $g_{mod}(z)$ requires taking the expectation of the modifier with respect to this Gaussian. For small δ, the mean of Xw given z is centered near $\hat{z} = X(X^TX)^{-1}X^Tz = Pz$ the projection of z into the column space of X, and the covariance is near $\delta^2 P$.

An expansion of g_{mod} for small δ may aid the effort in designing a suitable modifier g, so as to make $g_{mod}(z)$ approximately match a multiple of $f(z)$ minus a constant, so that the distribution of z_t approximately tracks $p_{\gamma_t}(z)$. Toward this aim consider the case that $g(z, u) = g(z)$ depends only on z so that $g_{mod}(z) = g(z) - E[\tilde{g}(u)|z]$, where $\tilde{g}(u)$ is near $E[g(z)|u]$. Taylor expansion of these conditional means, by the so-called δ method, yields for sufficiently smooth g, the following approximation, valid to order δ^4 in probability, for the function $g_{mod}(z)$

$$g(z) - g(\hat{z}) - \delta^2\left[\frac{1}{2}\operatorname{trace}(I + P)\nabla\nabla^T g(\hat{z}) - \nabla f(\hat{z}) \cdot \nabla g(\hat{z})\right]$$

We omit an analogous more elaborate expression that arises for the case that $g(z, u)$ depends also on $u = Xw$.

There is potential to use the given expansion in two ways. One is to approximate the main terms $g(z) - g(\hat{z})$ as $(z - \hat{z})^T \nabla g(z)$ valid to accuracy of order δ^2 in probability, and seek choice of g such that $(z - \hat{z})^T \nabla g(z)$ matches $f(z)$, ignoring the order δ^2 terms. The second way to use the expression is to arrange to cancel the $g(z) - g(\hat{z})$. Indeed, suppose $g(z)$ depends on z only through $\hat{z} = Pz$. That is, it takes the form $r(Pz)$ for some function r. In this case, since P is a projection, evaluating it at either z or $\hat{z} = Pz$ leaves it unchanged and hence the leading term difference $g(z) - g(\hat{z})$ is 0. Then the order δ^2 expression becomes the dominant part of the expansion for g_{mod} and the task becomes to find r for which it is proportional to our target. For example we can consider the case that r takes the additive form $r(v) = \sum_{i=1}^n r_i(v_i)$ for v in R^n which permits further simplification of our expression for $g_{mod}(z)$. It reduces to

$$g_{mod}(z) = -\delta^2\left[\sum_{i=1}^n P_{ii} r_i''(\hat{z}_i) - \gamma \sum_{i=1}^n (P\nabla f(\hat{z}))_i r_i'(\hat{z})\right].$$

This comes close to permitting a term by term solution. Suppose in particular that the design X is such that the projection entries P_{ii} are all strictly positive. Then we may set the univariate function r_i to be a twice integrated $-[f_i(z_i) - m_i]/P_{ii}$, that is,

$$r_i(v_i) = -\tfrac{1}{P_{ii}} \int_{a_i}^{v_i} \int_{a_i}^{\tau} [f_i(t) - m_i] dt d\tau$$

where a_i is a convenient reference point, such as 0 or a point where $f_i(a_i) = m_i$. Here m_i is the mean of $f_i(z_i)$ with respect to p_γ at the current gain γ_t, which can be estimated by an average across multiple parallel chains at the current t.

With this choice for the r_i the first term above matches what we desire and the second term remains. We end the story for now by calling attention to the following circumstance. If the objective function $f(z)$ has a gradient with zero projection $P\nabla f(z) = 0$ into the span of X, this second term vanishes and $g_{mod}(z)$ matches what we want in order to track the distribution. This encourages reformulation of the problem such that the linear part of the objective function is removed. Such reformulation may be natural for estimating ridge functions. We encourage its further consideration.

In conclusion we have examined the behavior of Markov chains with transitions based on specified drifts plus noise and by Gibbs samplers. Evolution of the density in a desired manner is shown to require a modification of transition rules previously dictated by invariance. In particular we have given preliminary analysis of modification of the drift in the approximate diffusion case and a multiplicative modification of the transition density in the Gibbs sampler case. The ideal modifying functions are solutions to partial differential equations which we have exhibited here. The hope is to encourage further research in this path to better understand Markov Chain evolution, and to determine which situations permit practical optimization.

References

[1] D. S. Johnson and F. P. Preparata, “The densest hemisphere problem,” *Theor. Comp. Sci.*, vol. 6, pp. 93–107, 1978.

[2] R. Greer, *Trees and Hills: Methodology for Maximizing Functions of Systems of Linear Relations*, P. L. Hammer, Ed. North-Holland, 1984.

[3] A. Blum and R. Rivest, “Training a 3-node neural network is np-complete,” *Neural Networks*, vol. 5, pp. 117–127, 1992.

[4] B. DasGupta, H. Siegelmann, and E. Sontag, *On the Intractability of Loading Neural Networks*. Kluwer Academic Publishers, 1994, ch. X, pp. 357–389.

[5] V. Vu, “On the infeasibility of training neural networks with small mean-squared error,” *IEEE Trans. on Information Theory*, vol. 44, no. 7, pp. 2892–1900, 1998.

[6] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, “Equations of state calculations by fast computing machines,” *Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[7] W. Hasting, “Monte carlo sampling methods using markov chains and their applications,” *Biometrika*, vol. 57, pp. 97–109, 1970.

[8] S. Geman and D. Geman, “Stochastic relaxation, gibbs distributions, and the bayesian restoration of images,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721–741, 1984.

[9] M. Tanner and W. Wong, “The calculation of posterior distributions by data augmentation (with discussion),” *Journal of Amer. Stat. Assoc.*, vol. 82, pp. 528–550, 1987.

[10] S. Kirkpatrick, C. D. G. Jr., and M. P. Vecchi, “Optimization by simulated annealing,” *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[11] J. Liu and R. Chen, “Sequential monte carlo methods for dynamic systems,” *Journal of the American Statistical Association*, vol. 93, no. 443, pp. 1032–1044, 1998.

[12] Y. Iba, “Population monte carlo algorithms,” *Transactions of the Japanese Society for Artificial Intelligence*, vol. 16, p. 279, 2000.

[13] D. Talay, *Simulation of Stochastic Differential Systems*. Springer-Verlag, 1995, ch. 3, pp. 63–106.

[14] J. Chang, “Stochastic processes, http://pantheon.yale.edu/ jtc5/251/ stochasticprocesses.pdf.”

[15] M. Dyer, A. Frieze, and R. Kannan, “Random polynomial time algorithm for estimating volumes of convex bodies,” *Proc 21st Annu ACM Symp Theory Comput*, pp. 375–381, 1989.

[16] A. Frieze and R. Kannan, “Log-sobolev inequalities and sampling from log-concave distributions,” *Annals of Applied Probability*, vol. 9, no. 1, pp. 14–26, 1999.

[17] L. Lovasz and S. Vempala, “The geometry of logconcave functions and sampling algorithms,” *Random Structures & Algorithms*, vol. 30, no. 3, pp. 307–358, 2007.

[18] P. Diaconis and D. Strook, “Geometric bounds for eigenvalues of markov chains,” *The Annals of Applied Probability*, vol. 1, no. 1, pp. 36–61, 1991.

[19] J. Fill, “Eigenvalue bounds on convergence to stationarity for nonreversible markov chains, with an application to the exclusion process,” *The Annals of Applied Probability*, vol. 1, no. 1, pp. 62–87, 1991.

[20] B. Hajek, “Cooling schedules for optimal annealing,” *Mathematics of Operations Research*, vol. 13, no. 2, pp. 311–329, 1988.

[21] A. R. Barron, “Universal approximation bounds for superpositions of a sigmoidal function,” *IEEE Transactions on Information Theory*, vol. 39, pp. 930–945, 1993.

[22] ——, “Approximation and estimation bounds for artificial neural networks,” *Machine Learning*, vol. 14, pp. 113–143, 1994.

[23] E. Candes, “Ridgelets: Estimating with ridge functions,” *The Annals of Statistics*, vol. 31, no. 5, pp. 1561–1599, 2003.

[24] L. K. Jones, “A simple lemma on greedy approximation in hilbert space and convergence rates for projection pursuit regression and neural network training,” *The Annals of Statistics*, vol. 20, no. 1, pp. 608–613, 1992.

[25] W. S. Lee, P. L. Bartlett, and R. C. Williamson, “Efficient agnostic learning of neural networks with bounded fan-in,” *IEEE Transactions of Information Theory*, vol. 42, no. 6, pp. 2118–2132, 1996.

[26] A. Barron, A. Cohen, W. Dahmen, and R. DeVore, “Approximation and learning by greedy algorithms,” *Annals of Statistics*, vol. 35, 2007.

[27] H. Risken, *The Fokker-Planck Equation: Methods of Solutions and Applications*. Springer, 1996.