Generalized Darting Monte Carlo

Cristian Sminchisescu
Toyota Technological Institute
Chicago, USA
crismin@nagoya.uchicago.edu

Max Welling
Bren School of Information and Computer Science
UC Irvine, CA 92697-3425, USA
welling@ics.uci.edu

Abstract

One of the main shortcomings of Markov chain Monte Carlo samplers is their inability to mix between modes of the target distribution. In this paper we show that advance knowledge of the location of these modes can be incorporated into the MCMC sampler by introducing mode-hopping moves that satisfy detailed balance. The proposed sampling algorithm explores local mode structure through local MCMC moves (*e.g.* diffusion or Hybrid Monte Carlo) but in addition also represents the relative strengths of the different modes correctly using a set of global moves. This 'mode-hopping' MCMC sampler can be viewed as a generalization of the darting method [1]. We illustrate the method on a 'real world' vision application of inferring 3-D human body pose from single 2-D images.

1 Introduction

It is well known that MCMC samplers have difficulty in mixing from one mode to the other because it typically takes many steps of very low probability to make the trip [2, 3]. Recent improvements designed to combat random walk behavior, like Hybrid Monte Carlo and over-relaxation [4, 2] do not solve this problem when modes are separated by high energy barriers. In this paper we show how to exploit knowledge of the location of the modes to design a MCMC sampler that mixes properly between them.

One can imagine at least two possible scenarios where this advance knowledge is present. In one scenario we are given data-cases and aim at learning a model distribution. In this case, the data itself is representative of the high probability regions of a well fitted model. To identify these regions one could first fit a mixture-of-Gaussians model and use the "one-standard deviation" ellipsoids of the fitted covariances as jump regions. These regions may overlap, but as we will see in the following, our algorithm is well suited to deal with this complication. This idea was explored in the context of learning Markov random field models using contrastive divergence learning with an alternative (we believe less efficient) mode-hopping algorithm [11]. There it was shown that mode-hopping can be of practical value in learning the correct relative volumes of the modes in the probability density.

In another scenario we have actively searched for high probability regions using sophisticated optimization methods [6, 7]. Given these local maxima, we are then interested in collecting unbiased samples from the underlying probability distribution. In this paper we explore the second scenario in the context of a "real world" computer vision application – learning human models and estimating high-dimensional 3D human body poses from 2D image information.

The main contribution of this paper is the introduction of a new mode-hopping sampler, the generalized darter, and its extensions. A proof of detailed balance is given in the appendix. Further details, including an auxiliary variable formulation and extensions to discrete state spaces can be found in an accompanying technical report [12].

2 Markov Chain Monte Carlo

Consider a given probability distribution $p(\mathbf{x})$ with $\mathbf{x} \in \mathcal{X} \subset R^d$ a vector of continuous random variables. In the following we will focus on continuous variables, but the algorithm is easily extended to discrete state spaces. A very general method to sample from this distribution is provided by Markov chain Monte Carlo (MCMC) sampling. The idea is to start with an initial distribution $p_0(\mathbf{x})$ and design a set of transition probabilities that will eventually converge to the target distribution $p(\mathbf{x})$.

The most commonly known transition scheme is the one proposed in the Metroplis-Hastings (M-H) algorithm, where a target point is sampled from a possibly asymmetric conditional distribution $Q(\mathbf{x}_{t+1}|\mathbf{x}_t)$, where $\mathbf{x}_t$ represents the current sample. To make sure that detailed balance holds, *i.e.* $p(\mathbf{x}_t)Q(\mathbf{x}_{t+1}|\mathbf{x}_t) = p(\mathbf{x}_{t+1})Q(\mathbf{x}_t|\mathbf{x}_{t+1})$, which in turn guarantees that the target distribution remains invariant under Q, we should only accept a certain fraction of the proposed targets. In the most commonly used M-H algorithm, the transition distribution Q is symmetric and independent of the energy-surface at location $\mathbf{x}$. This simplifies the acceptance equation (the Q factors cancel), but leads to slow mixing due to random walk behavior. It is however not hard to incorporate local gradient information, $d\log p(\mathbf{x})/d\mathbf{x}$ to improve mixing. One could for instance bias the proposal distribution $Q(\mathbf{x}_{t+1}|\mathbf{x}_t)$ in the direction of the gradient $d\log p(\mathbf{x})/d\mathbf{x}$. When the stepsize becomes infinitesimally small this is called the Langevin method and one can show that the rejection rate vanishes in this limit. This strategy improves the acceptance rate, but mixing between different modes remains difficult.

3 Mode-Hopping MCMC

We start with reviewing the closely related darting algorithm described in [1]. In darting-MCMC we place spherical jump regions of equal volume at the location of the modes of the target distribution. The algorithm is based on a simple local MCMC sampler which is interrupted with a certain probability to check if its current location is inside one of these spheres. If so, we initiate a jump to the corresponding location in another sphere, chosen uniformly at random, where the usual Metropolis acceptance rule applies. To maintain detailed balance we decide not to move if we are located outside any of the balls. It is not hard to check that this algorithm maintains detailed balance between any two points in sampling space.

In high dimensional spaces this procedure may still lead to unacceptably high rejection rates because the modes will likely decay sharply in at least a few directions. Since these ridges of probability are likely to be uncorrelated across the modes, the proposed target location of the jump will have very low probability, resulting in almost certain rejection. In the following we will propose two important improvements over the darting method. Firstly, we allow the jump regions to have arbitrary shapes and volumes and secondly these regions may overlap. The first extension opens the possibility to align the jump regions precisely with the shape of the high probability regions of the target distribution. The second extension simplifies the design and placement of the jump regions since we don't have to worry about possible overlaps of the chosen regions.

First consider the case when the regions are non-overlapping but of different volumes. Like in the darting method we could consider a one-to-one mapping between points in the different regions, or we could choose to sample the target point uniformly inside the new region. Because the latter is somewhat simpler conceptually, we'll use uniform sampling in this section. The deterministic case will be treated in the next section. Also, to simplify the discussion we'll first consider the case where the underlying target distribution is uniform, *i.e.* has equal probability everywhere. Due to the difference in volumes, particles are more likely to be inside a large region than in small ones. Thus, there will be a larger flow of particles going from the bigger regions towards the smaller ones violating detailed balance. To correct for it we could reject a fraction of the proposed jumps from larger towards smaller regions. There is however a smarter solution, that picks the target region proportional to its volume:

$$P_i = \frac{V_i}{\sum_j V_j} \tag{1}$$

If we view the jumps between the various regions as a (separate) Markov chain, this method samples directly from the equilibrium distribution while a rejection method would require a certain mixing time to reach equilibrium. Clearly, if the underlying distribution is not uniform, we need the Metropolis acceptance rule between the jump point and its image in the target region:

$$P_{accept} = \min\left[1, \frac{p(\mathbf{t})}{p(\mathbf{x})}\right] \tag{2}$$

where $\mathbf{t}$ is the target point and $\mathbf{x}$ is the exit point.

Now, let's see what happens if two regions happen to overlap. Again, we first consider sampling the target point uniformly in the new region, and consider a target distribution which is uniform. Consider two regions which partly overlap. Due to the fact that we use the probability P_i (1), each volume element $\mathbf{dx}$ inside the regions has equal probability of being chosen. However, points located in the intersection will be a target twice as often as points outside the intersection. To compensate, *i.e.* to maintain detailed balance, we need to reject half of the proposed jumps into the intersection. In general, we check the number of regions that contain the exit point, $n(\mathbf{x})$, and similarly for the target point, $n(\mathbf{t})$. The appropriate fraction of moves that is to be accepted in order to maintain detailed balance is $\min[1, n(\mathbf{x})/n(\mathbf{t})]$. Combining this with the Metropolis acceptance probability 2 we find:

$$P_{accept} = \min\left[1, \frac{n(\mathbf{x})p(\mathbf{t})}{n(\mathbf{t})p(\mathbf{x})}\right] \tag{3}$$

Generalized Darting MCMC Sampler

Repeat until convergence

1. Draw a sample $u_1 \sim U[0,1]$.
2. if $u_1 > P_{check}$: perform one step of a local MCMC sampler.
3. if $u_1 < P_{check}$
   (a) Identify the number of regions $n(\mathbf{x})$ that contain the current sample.
   (b) if $n(\mathbf{x}) = 0$ do nothing.
   (c) if $n(\mathbf{x}) > 0$
      i. Sample a new region according to P_i (1).
      ii. Propose a location inside the new region (either deterministically or uniformly at random).
      iii. Identify the number of regions $n(\mathbf{t})$ that contain the proposed sample.
      iv. Draw a sample $u_2 \sim U[0,1]$.
      v. if $u_2 > P_{accept}$ (3) reject move.
      vi. if $u_2 < P_{accept}$ (3) accept move

Figure 1: The steps of our generalized darting sampler.

Putting everything together, we define the mode-hopping MCMC sampler explained in figure 1.

3.1 Elliptical Regions & Deterministic Moves

In the previous section we have uniformly sampled the proposed new location of the particle inside the target region. This is a very flexible method for which it is easy to prove detailed balance. However, a deterministic transformation can be tuned to map between points of roughly equal probability which is expected to improve the acceptance rate. Consider for instance the case that the energy surfaces near the regions is exactly quadratic and have the same height (*i.e.* their centers have equal probability). We can now define a transformation between ellipses that maps between points of equal probability resulting in a vanishing rejection rate. This is obviously not the case when we use uniform sampling.

We first consider the case of non-overlapping elliptical regions. Ellipses seem a natural choice, but the algorithm presented here is by no means restricted to it. For instance, the method is readily generalized to the use of rectangles as basic shapes. We'll parameterize an ellipse by a mean $\boldsymbol{\mu}$, a covariance $\boldsymbol{\Sigma}$ and a scale α, *i.e.* the ellipse is defined to be the equiprobability contour that is α standard deviations away from the mean. We will also need the eigenvalue decomposition of the covariance, $\boldsymbol{\Sigma} = \mathbf{USU}^\top$, where $\mathbf{S}$ is a diagonal matrix containing the eigenvalues $\{\lambda_i\}$. A deterministic transformation between two ellipses $1 \to 2$ is given by:

$$\mathbf{x}_2 = \boldsymbol{\mu}_2 - \mathbf{U}_2\mathbf{S}_2^{1/2}\mathbf{S}_1^{-1/2}\mathbf{U}_1^\top(\mathbf{x}_1 - \boldsymbol{\mu}_1) \qquad (4)$$

We note that this transformation would not leave a point invariant if we chose the second ellipse to be equal the first one, but mirrors it in the origin. Even though the transformation above is one-to-one, it does change the volume element $\mathbf{dx}$, implying that we need to take the Jacobian of the transformation into consideration. The intuitive reason for this is the same as in the previous section: more particles will be located in the larger ellipses resulting in more jumps to smaller ellipses than back, violating detailed balance. To compensate we sample the target ellipse again proportional to its volume ($\alpha^d \prod_{i=1}^d \lambda_i$), *i.e.* using (1).

We will now discuss how this algorithm can be generalized in case the ellipses overlap. Consider again two ellipses which partly overlap and a uniform target density. Consider a point that is located inside both ellipses, *i.e.* in the overlap (point 1). To apply the deterministic mapping, we first need to choose one of the two ellipses as a basis for the transformation. Unfortunately, an arbitrary rule such as the ellipse on top of the stack, or the one with the largest volume will result in a violation of detailed balance. Thus, we propose to pick the ellipse at random with equal probability. Now consider the image point under the mapping (point 2), choosing either the same ellipse (resulting in mirroring the point at the origin) or choosing the other ellipse. Assume point 2 is not located in the overlap. The probability of moving from $1 \to 2$ is $\frac{1}{4}$; a factor $\frac{1}{2}$ coming from the fact that we first choose with equal probability which ellipse will be used to define the transformation, and another factor $\frac{1}{2}$ because we sample the target ellipse using (1). However, in the other direction $2 \to 1$ the probability is $\frac{1}{2}$. Note that unlike the case of uniformly sampling a target point (see previous section) the probability of going from $2 \to 1$ is not doubled[1]. Thus, to rescue detailed bal-

[1]The reason is that for every target ellipse the image of the point under the mapping (4) is different. However, there are circumstances, e.g. when one ellipse is completely encircled by a larger one, that isolated points have the same image for two distinct target ellipses, resulting in violation of detailed balance. Since in the continuous case this set

ance we need to accept only half of the proposed moves from $2 \rightarrow 1$, or more generally $\min[1, n(\mathbf{x})/n(\mathbf{t})]$ with $n(\cdot)$ the number of ellipses containing a point. Combining this with the usual Metropolis acceptance rule applicable to general target densities, we arrive precisely at the rule in (3).

To summarize, the deterministic algorithm has precisely the same structure as algorithm in fig. 1, where in the transformation (4) ellipse 1 is chosen uniformly at random from all ellipses containing point 1 and ellipse 2 is chosen using (1) with V_i the volume of the ellipsoid used to approximate the region.

4 Monocular Human Pose Estimation

We explore the potential of the generalized darting method for *monocular* 3D human pose estimation. This problem has applications for human-computer interaction and for actor reconstruction from movie footage – in this case only one camera viewpoint, the one presented in the movie, is usually available.

We run experiments based on correspondences between the articulated joints of a subject in the image and the joints of a 3D articulated model (2D-3D correspondences). We also report experiments for learning the model parameters in a maximum likelihood framework using a more sophisticated edge-based observation model. Monocular human pose estimation is well adapted to illustrate the algorithm because the resulting 3D pose posterior is both high-dimensional ($\approx$35 human joint angle state variables) and highly multimodal. In any single monocular image, under point-wise 3D human joints and their image projections, each limb of the human is subject to a 'reflective' kinematic flip ambiguity. Two 3D human body configurations with symmetrical slant in depth w.r.t. the camera (see fig. 4) produce identical point-wise image perspective projections. The number of possible solutions multiples over the number of links of the human body. For example, a 3D human model with 10 links (torso, head, left/right forearm, upperarm, thigh and calf) may have $2^{\#links}$ local optima, although this is usually an overestimate. Some solutions may not be physically plausible and may violate joint angle limits or body non-self-intersection constraints. The question this work addresses is not how to find the optima but how to efficiently sample from the 3D human pose *equilibrium distribution* once these are known.

4.1 Domain Modeling

This section describes the humanoid visual models used in our sampling experiments. For details see [8].

has measure zero, we will ignore it.

Representation: A typical human body model is constructed from a 'skeleton' that has 30-35 rotational joints controlled by angular joint state variables $\mathbf{x}$, which includes a global 6d translation of the body center. It also has 'body flesh' built from three-dimensional ellipsoids with deformation parameters $\boldsymbol{\theta}$, here 36 variables for the head, torso, arms and legs. The surface model improves the image representation for the 3D human pose estimates based on image features like edges. In one of the experiments we not only estimate the model state, but also learn its parameters using maximum likelihood.

Joint positions $\mathbf{u}_i$ in local coordinate systems for each body limb are transformed into points $\mathbf{p}_i(\mathbf{x}, \mathbf{u}_i)$ in a global 3-D coordinate system, then into predicted image points $\mathbf{r}_i(\mathbf{x}, \mathbf{u}_i)$ using composite nonlinear transformations $\mathbf{r}_i(\mathbf{x}, \mathbf{u}_i) = \mathbf{P}(\mathbf{p}_i(\mathbf{x}, \mathbf{u}_i)) = \mathbf{P}(\mathbf{K}(\mathbf{x}, \mathbf{u}_i))$, where $\mathbf{K}$ represents a chain of rigid transformations that map different body links through the kinematic chain to their global 3-D position (see fig. 2), and $\mathbf{P}$ represents perspective image projection.

Observation Likelihood: For model state estimation, we compute the negative log likelihood of each known image joint position, $\mathbf{o}_i$, under a Gaussian centered at its projected (hypothesized) image location, $\mathbf{r}_i$. The costs are summed over all the human body joints to produce the state space energy function. This is a function $e(\mathbf{o}_i|\mathbf{x}, \boldsymbol{\theta})$ of the prediction error $\Delta\mathbf{o}_i(\mathbf{x}) = \mathbf{o}_i - \mathbf{r}_i(\mathbf{x})$ between the model and the given image data, $e(\mathbf{o}_i, \mathbf{x}, \boldsymbol{\theta}) = \Delta\mathbf{o}_i^2/2\sigma^2$. The cost gradient $\mathbf{g}_i(\mathbf{x})$ and Hessian $\mathbf{H}_i(\mathbf{x})$ are also computed, being used for second order continuous optimization and hybrid Monte Carlo (HMC) step calculations. For learning experiments, we use an observation model based on edge residuals. These are collected at model occluding contours predicted in the image. At each 3D model configuration, for each element on an image-predicted model contour, a line search along the normal direction is used to locate an image edge that matches it. The distance between the location of the model contour and the image edge is used to construct a quadratic function of the residual, similar to the one based on skeletal joint residuals. This is summed over all contour predictions and all the human body parts.

Energy Function: The model state estimates are obtained by optimizing a maximum *a posteriori* criterion, the total posterior probability according to Bayes rule:

$$\begin{aligned} p(\mathbf{x}|\mathbf{O}, \boldsymbol{\theta}) &\propto p(\mathbf{O}|\mathbf{x}, \boldsymbol{\theta})\, p(\mathbf{x}) \\ &\propto \exp\left(-\textstyle\sum_i e(\mathbf{o}_i, \mathbf{x}, \boldsymbol{\theta})\right) p(\mathbf{x}) \end{aligned} \tag{5}$$

where $e(\mathbf{o}_i, \mathbf{x})$ is the cost density associated with observation i, the sum is over all observations $\mathbf{O} = \{\mathbf{o}_1, ..., \mathbf{o}_n\}$, and $p(\mathbf{x})$ is the model state prior. For experiments, we have used the classical Langevin sampler

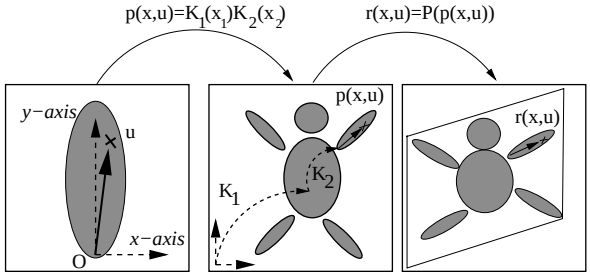

Figure 2: A simple model of a kinematic chain consisting of ellipsoidal parts. First, the feature $\mathbf{u}_i$, defined in its local coordinate frame, is mapped to a 3-D position, $\mathbf{p}_i(\mathbf{x}, \mathbf{u}_i)$, in the body model through a chain of transformations, $\mathbf{K}_i(\mathbf{x}_i)$, between local coordinate systems. $\mathbf{x}_i$ are state variables that encode transformations (rotation angles) between these reference frames, collectively stored in a vector $\mathbf{x}$. The 3-d surface point given by $\mathbf{p}_i(\mathbf{x}, \mathbf{u}_i)$ is mapped to the image using perspective projection: $\mathbf{r}_i(\mathbf{x}, \mathbf{u}_i) = \mathbf{P}(\mathbf{p}_i(\mathbf{x}, \mathbf{u}_i))$, where $\mathbf{P}$ is the viewing camera projection matrix (this includes the global orientation of the camera and its intrinsic parameters, *e.g.* focal length, pixel dimensions, *etc.*).

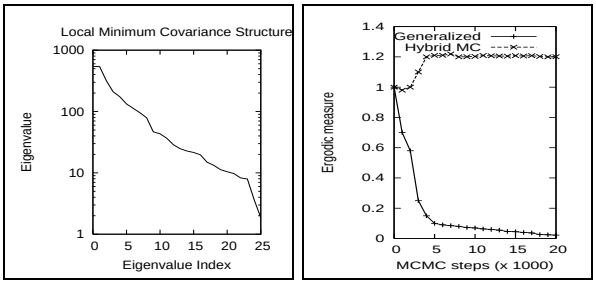

Figure 3: (a, left) Top 25 eigenvalues of the covariance matrix (corresponding to a 35 state variable model) for a local minimum shows the typical ill-conditioning of the monocular human pose estimation. (b, right) The ergodic measure compared for a classical Langevin gradient-based sampling scheme and the generalized darting method. The classical sampler doesn't mix well – the long-term energy difference between trajectories reflects the memory of the minima where they were initiated. The generalized method mixes much better and explores various minima so the average state-space difference over long-term trajectories tends to zero (see text).

(see section 2), in combination with long-range jumps using the spherical darting and the generalized darting methods.

Prior Distributions: The priors we use are stabilizers to avoid singular distributions for hard to estimate state variables (*e.g.* in the clavicle and shoulder complex), and terms for collision avoidance between body parts, and joint angle limits.

4.2 Experiments

Sampling Experiments: We have selected 4 local minima corresponding to the left forearm and left calf in a monocular side view of the body (see fig. 4). The local minima have relative volumes of $(0.16, 0.38, 0.10, 0.36)$ and energy levels $(4.41, 6.31, 7.20, 8.29)$.

For local optimization we use a second-order damped Newton trust region method [10] where gradient and Hessians of the energy functions are computed analytically and assembled using the chain rule with back-propagation on individual kinematic chains. For generalized darting, we estimate local covariances as inverse Hessians at each local minimum. For MCMC simulations, we enforce joint limit constraints using reflective boundary conditions, *i.e.* by reversing the sign of the normal momentum when it hits a joint limit. We found this gave an improved sampling acceptance rate compared to simply projecting the proposed configuration back on the constraint surface, as the latter leads to cascades of rejected moves until the momentum direction gradually swings around.

We ran the simulation with stepsize $\Delta\tau = 0.1$, using the Langevin sampler (fig. 5), the darting method with spherical covariances (fig. 6a) and the generalized darting method with deterministic moves (fig. 6b). In fig. 6 we show a fragment of a larger simulation that uses a small jump probability $P = 0.03$, in order to diminish the frequency of jumps for illustrative purposes. It is easily noticeable that the classical sampler is trapped in the starting mode, and wastes all of its samples exploring it repeatedly. The spherical darting method explores only 2 minima based on one successful long-range jump during 600 iterations. The darting method (right) explores more minima by combining local moves with non-local jumps that are accepted more frequently. Different minima are visited using 7 jumps. This could be visually observed in fig. 6. After each jump, the sampler equilibrates at a different energy level associated to the new local minimum.

We have also performed a large simulation (10^5 steps) with $\Delta\tau = 0.1$ and probability $P = 0.25$ for the darting moves. The first 200 samples were discarded in order to let the chain reach equilibrium. The covariance volume scaling factor α was set to unity for all regions. For classical darting, we place spheres of unit radius around each minimum. With these parameters, the sampler mixes fast within each minimum, but still has good acceptance rates of 94% for local moves. The acceptance rate for long-range jumps in the spherical case is $a_s = 1292/24,863 = 0.052$ whereas for the generalized darting case is $a_g = 9642/25,850 = 0.388$, which is an important improvement. According to our tests, the results are stable to changes in the volume factor α by roughly 10%.

Ergodicity Study: To show the benefit of incorpo-

rating mode-hopping moves into the MCMC sampler we compared the performance of a generalized darter and a hybrid MCMC sampler in an experiment based on 3 runs of $20,000$ simulation steps each, following the same paradigm as in the previous section. We compute the ergodic measure [1], an indicator for the rate of self-averaging in equilibrium calculations. Although self-averaging is a necessary but not sufficient condition for the ergodic hypothesis to be satisfied, it gives intuition about the rate of state space sampling. We have selected the state-space configuration as the quantity to average (alternatively an ergodic measure based on some other property, *e.g.* the energy could be used). This measure is an average over pairwise differences between average state-space positions, for trajectories initiated in different minima during a simulation. More specifically, the average state space position after S moves from a trajectory *initiated* at minimum a, containing configurations $\{\mathbf{x}_i^a, i = 1..S\}$ obtained[2] during sampling run k is given by:

$$d_k^a(S) = \frac{1}{S}\sum_{i=1}^{S} ||\mathbf{x}_i^a|| \tag{6}$$

and the ergodic measure is defined as the average between two trajectories initiated at different minima a and b in R runs[3]:

$$e(a, b, S, R) = \frac{1}{R}\sum_{k=1}^{R}[d_k^a(S) - d_k^b(S)]^2 \tag{7}$$

For good mixing over large trajectories we expect the ergodic measure to converge to 0. In fig. 3, we plot the ergodic measure corresponding to a classical Langevin simulation with no jumps against one using the generalized darting scheme for $S = 20,000$ over $R = 3$ runs. The mixing of the classical hybrid MCMC sampler is not satisfactory, perhaps reflecting the average state-space difference between the two local minima where the sampler is trapped, and which are explored repeatedly. In contrast, the long-range state self-averaging effect is clearly observed for generalized darting.

Learning Experiments: We run a parameter learning experiment using the same image in fig. 4, the same state priors but using a more complex image *observation likelihood* based on *contour / edge measurements* (see §4.1). We estimate some of the model parameters, here the body proportions (36 parameters representing the superquadrics of the head, torso, upperarm, forearm, thigh and calf, with the symmetrical values on the left and right side of the body mirrored) and the variance used for the quadratic edge residual cost.

[2]The trajectory may well include configurations inside minima basins other than a, but in a slight abuse of notation we will identify *both* the starting minima *and* the trajectory itself with the same letter.

[3]Note that there are two *different* simulations for each run k, one for a and another for b. Also notice that the subscript does not index the vector $\mathbf{x}$ but indicates different state vectors.

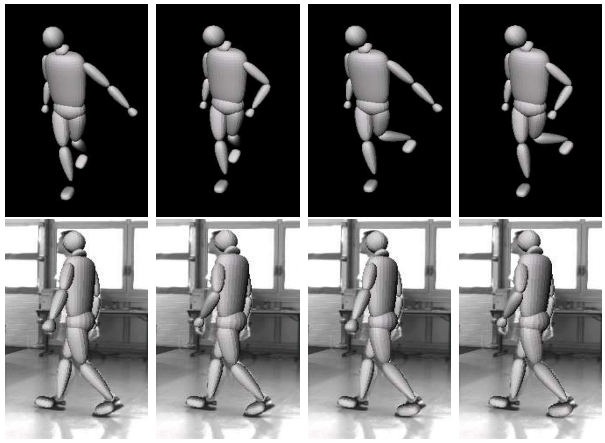

Figure 4: Human pose estimation based on a single image of a walking person photographed sideways. In any single monocular image, under point-wise 3D human joints and their image projections, each limb of the human is subject to a 'reflective' kinematic flip ambiguity. Two 3D human body configurations with symmetrical slant in depth w.r.t. the camera produce identical point-wise image perspective projections. The bottom row shows four copies of the same image, with the projection of four different poses of the model superimposed on the image. The four poses are shown from a different viewpoint in the top row. The different poses correspond to four local minima in the energy function, defined over 35 state variables (the human joint angles). Notice how the four human body configurations indeed align well with the imaged human.

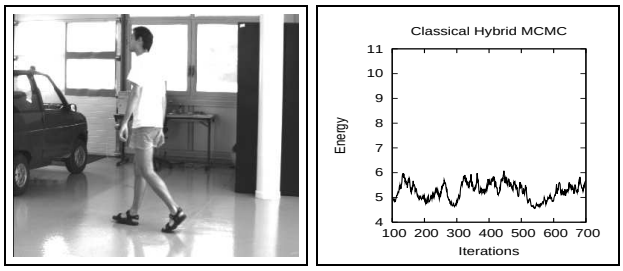

Figure 5: (left) Original image used for human pose inference and learning. (right) Classical hybrid Monte Carlo gets trapped in the starting minimum.

We learn by maximizing the conditional Maximum Likelihood using gradient ascent. To obtain estimates for the gradients we need to compute the gradient of the log-normalization term associated with $p(\mathbf{O}|\mathbf{x}, \boldsymbol{\theta})$ in equation 5. This derivative is intractable but can be approximated by drawing samples from the model using the generalized darting MCMC algorithm. This procedure is supervised, *i.e.* we need to specify the 3D state ground truth. Since this information was not available for our real scene, we selected, by visual inspection, one of the 4 pose configurations, considered to be the most plausible, as the ground truth

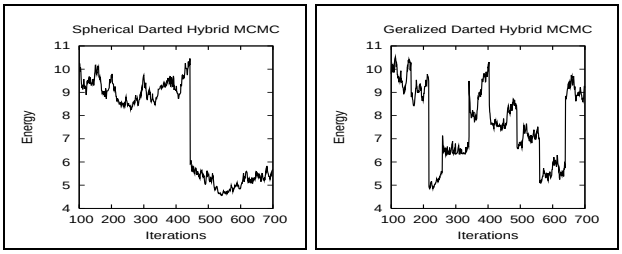

Figure 6: Spherical darting (a,left) explores the minima more thoroughly, but only 2 minima are visited during 600 iterations (1 successful jump). Generalized darting (b,right) explores the different minima by combining local moves with non-local jumps that are accepted frequently. 8 local minima are visited via 7 jumps (after each jump the sampler explores its new local minimum for a while before the next jump).

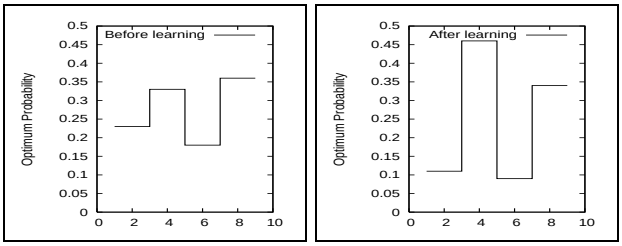

Figure 7: Learning the body proportions and the variance of the observation likelihood based on matching contours improves the relative probability of the correct solution w.r.t. to undesirable ones. The plots show the probability levels of four different local optima (the numbering on the horizontal axis is irrelevant). The order of the probability peaks corresponds to the one shown in fig. 4, with the second configuration visually selected as ground truth. Learning significantly increases the probability of the desired solution and downgrades competing, incorrect ones.

(the second column in fig. 4). The probability of the 4 configurations *before* and *after learning* is shown in fig. 7. Notice how the process substantially improves the margin between the correct solution and the undesirable ones, in particular how the selected ground-truth emerged as most probable *after* learning despite not being the most probable *before*. 3D pose estimates based on the learned model identify the correct solution with significantly higher probability, on the average. Learning does not make all the incorrect solutions extremely implausible due to several reasons. First, there is 3D structure sharing between the incorrect and the 'ground-truth' (*e.g.* solutions 2 and 4 share the upper body sub-component of the state). Another factor may be the weak evidence provided by the contour features used for the observation likelihood. One can, *e.g.* use better lighting or surface reflection models in order to provide additional constraints to diminish uncertainty. Finally, since we are only able to find a local optimum for the parameters, it is possible that other good ones exist. However, we haven't empirically identified better ones, even after multiple restarts from different initial starting points.

5 Discussion

In this paper we have discussed a new Markov chain Monte Carlo sampler, that is able to effectively jump between modes in the target distribution while maintaining detailed balance. Our method is a generalization of 'darting MCMC' where the basic jump regions may have an arbitrary irregular shape and moreover are allowed to overlap. We demonstrate the algorithms on model learning and inference for 3D human body pose estimation from monocular images.

Generalized darting is easily extended to discrete state spaces. Regions can be defined in any suitable manner, e.g. by a set of conditions that should hold, as long as we can count the number of states in each region. Overlaps between regions are also allowed. The algorithm proposes to jump into a region with probability proportional to the number of states in that region, picking a state uniformly at random, and accepting this move again according to equation 3.

An alternative view of the proposed generalized darting sampler is that of a mixture between an independence sampler and a Hybrid Monte Carlo sampler. In this view, we randomly alternate HMC sampling with proposing samples uniformly from the collection of regions $\{V_i\}$. The proposal distribution is not conditional on the previous sample, hence the name "independence sampler". However, to maintain detailed balance we can not accept a proposal if the previous sample was located outside this collection of regions. Hence, instead of performing this check before proposing a new sample (as in darting MCMC), the check is implicitly performed after proposing a new sample by incorporating it in the acceptance rule.

Apart from the darting method, other MCMC schemes that mix between distant modes can be found in the literature (e.g. 'simulated tempering' [5], normal kernel coupling [14] and a method that incorporates deterministic optimization [13]). The main advantage of the proposed generalized darting method is that one can tune the shape of the jump regions to match the shape of the high probability regions of the target distributions. This should help to achieve an improved acceptance probability of attempted jumps between regions. Note however that we do not claim that our method is superior to all earlier schemes under all circumstances. In fact, we have only compared our method with the classical darting method and shown improved acceptance rates. No doubt, the various methods described above will have different properties for different target distributions, or in the presence of different amounts of prior knowledge about the target distribution.

A Proof of Detailed Balance

The generalized darting Monte Carlo sampler can be viewed as a Hybrid Monte Carlo sampler that is interrupted with a certain probability to attempt a long range jump. Since Hybrid Monte Carlo sampling is ergodic, a "mixture" of Hybrid Monte Carlo and any other (possibly non-ergodic) sampler is automatically ergodic as well. To prove detailed balance between any pair of points in the sample space, we consider the following three possibilities:

1: Both points are located in one or more of the regions. We will prove the case of two points in possibly overlapping regions, where the jump points are sampled uniformly at random inside a target region. The prove for the deterministic case goes along similar lines (see section 3.1). With probability $1 - P_{check}$ we follow the local dynamics of the Markov chain which fulfills detailed balance by assumption. With probability P_{check} we initiate a jump to some other point in some other region. Define A to be the set of regions that contain point 1 and B the set of regions that contain point 2. We now have:

$$
\begin{aligned}
& p(\mathbf{x}_1)\ P(\mathbf{x}_1 \to \mathbf{x}_2) \\
&= p(\mathbf{x}_1)\ P_{check}\ \sum_{i \subset B} \frac{P_i}{V_i}\ P_{accept:1 \to 2} \\
&= p(\mathbf{x}_1)\ P_{check}\ \frac{n(\mathbf{x}_2)}{\sum_j V_j}\ \min\left[1, \frac{p(\mathbf{x}_2) n(\mathbf{x}_1)}{p(\mathbf{x}_1) n(\mathbf{x}_2)}\right] \\
&= P_{check}\ \frac{1}{\sum_j V_j}\ \min\left[p(\mathbf{x}_1) n(\mathbf{x}_2), p(\mathbf{x}_2) n(\mathbf{x}_1)\right] \\
&= p(\mathbf{x}_2)\ P_{check}\ \frac{n(\mathbf{x}_1)}{\sum_j V_j}\ \min\left[1, \frac{p(\mathbf{x}_1) n(\mathbf{x}_2)}{p(\mathbf{x}_2) n(\mathbf{x}_1)}\right] \\
&= p(\mathbf{x}_2)\ P_{check}\ \sum_{i \subset A} \frac{P_i}{V_i}\ P_{accept:2 \to 1} \\
&= p(\mathbf{x}_2)\ P(\mathbf{x}_2 \to \mathbf{x}_1)
\end{aligned}
$$

where P_i (equation 1) is the probability of jumping to region i and the factor $1/V_i$ is included because the target point is sampled uniformly at random inside this region.

2: One of the two points is located inside one or more jump-regions. The particle located outside any jump-region follows its local dynamics with probability $1 - P_{check}$. The particle inside one or more regions will also follow its local dynamics with probability $1 - P_{check}$. With probability P_{check} the sampler decides to perform a check. But in that case the particle outside any region will stay put while the particle inside one or more regions will attempt a jump and will therefore never end up outside the set of all regions.

3: Both points are located outside any of the jump-regions. In this case, detailed balance follows because of the Markov chain for the local moves is assumed to respect detailed balance. With probability P_{check} this Markov chain is interrupted to check if the particle is located inside a jump-region. But since both points under consideration are assumed to be located outside any jump-region this interruption will be symmetric and does not destroy detailed balance.

Acknowledgments This material is based upon work supported by the National Science Foundation. CS was supported under grant IIS-0535140, MW under grant no. 0447903. We thank Geoff Hinton for feedback and earlier contribution to this work . We also thank Radford Neal and Yee-Whye Teh for general comments.

References

[1] I. Andricioaiei, J. Straub, A. Voter, Smart Darting MonteCarlo, J. Chem. Phys. (2001) 114 (16).

[2] R. Neal, Probabilistic Inference Using Markov Chain Monte Carlo, Tech. Rep. CRG-TR-93-1, University of Toronto (1993).

[3] G. Celeux, M. Hurn, C. Robert, Computational and inferential difficulties with mixture posterior distributions, J. Amer. Statist. Assoc 95 (2000) 957–979.

[4] S. Duane, A. D. Kennedy, B. J. Pendleton, D. Roweth, Hybrid Monte Carlo, Physics Letters B 195 (2) (1987) 216–222.

[5] Marinari, E. and G. Parisi: 1992, 'Simulated Tampering: A New Monte Carlo Scheme, Europhysics Letters **19**(6).

[6] R. Neal, Annealed Importance Sampling, Statistics and Computing 11 (2001) 125–139.

[7] C. Sminchisescu, B. Triggs, Building Roadmaps of Local Minima of Visual Models, in: European Conference on Computer Vision, Vol. 1, Copenhagen, 2002, pp. 566–582.

[8] C. Sminchisescu, B. Triggs, Estimating Articulated Human Motion with Covariance Scaled Sampling, International Journal of Robotics Research 22 (6) (2003) 371–393.

[9] C. Sminchisescu, M. Welling, Generalized Darting Monte-Carlo, Technical Report, University of Toronto, CSRG-543, October (2006).

[10] R. Fletcher, Practical Methods of Optimization, in: John Wiley, 1987.

[11] G. Hinton, M. Welling, A. Mnih, Wormholes Improve Contrastive Divergence, in: Advances in Neural Information Processing Systems, 2003.

[12] C. Sminchisescu, M. Welling, Generalized Darting Monte-Carlo, Tech. Rep. CSRG, University of Toronto (October 2006).

[13] Tjelmeland, H. and B. K. Hegstad: 1999, 'Mode Jumping Proposals in MCMC', Technical report, Norwegian University of Science and Technology, Trondheim, Norway, Preprint Statistics No.1/1999.

[14] Warnes, G.: 2000, The Normal Kernel Coupler: An adaptive Markov Chain Monte Carlo Method for Efficiently Sampling from Multi-modal Distributions, PhD thesis, University of Washington.