Stochastic Optimization of Black-box Functions on Riemannian Manifolds Using Kernel Adaptive SMC

Hanchen Xiong Sandor Szedmak Phillip Zech Justus Piater
Institute of Computer Science
University of Innsbruck
Technikerstr.21a, Innsbruck, A-6020, Austria
{*hanchen.xiong, sandor.szedmak, phillip.zech, justus.piater*}*@uibk.ac.at*

Abstract

This paper puts forward a new stochastic method for global optimization of black-box functions on a Riemannien manifold $\mathcal{M}$, which is rather challenging yet useful in many areas. The proposed algorithm is based on *Kernel Adaptive Sequential Monte Carlo* (KASMC) by exploiting the closeness between simulated annealing and sequential Monte Carlo (SMC), and thus is referred to as KASMC optimizer. At each Markov chain Monte Carlo (MCMC) transition of KASMC optimizer, samples are first proposed in a reproducing kernel Hilbert space (RKHS) of the embeding space of $\mathcal{M}$, after adaptively tempered acceptance criterion, they are mapped back to the embeding space and $\mathcal{M}$ respectively by two projection steps. The strength of KASMC optimizer, compared to classical simulated annealing, stems from its two adaptive tunning mechanisms. The first one is adaptive tempering, which automatically constructs temperature schedule through cooling; the second one is adaptive MCMC transition in RKHS, which is critical to fit highly nonlinear support of functions. Furthermore, a finite-time performance guarantee of KASMC optimizer is provided. Based on our experimental results on $SO(3)$ manifold and Stiefel manifold, KASMC demonstrates promising applicabilities.

1 Introduction

Global optimization on a black-box function f is a long-standing challenge task despite its wide usage and demand in many disciplines [28, 11]. Difficulties are usually considered from two aspects: *i.* the analytic form of f is unknown and function values can be only accessed point-wisely, thus no function landscape property (*e.g.* convexity or local convexity) can be exploited for numerical solutions; *ii.* the function can be arbitrarily complex, *e.g.* high-dimensional, multi-modal, which tends to trap solutions in local optimum. In this paper, we consider an even more "nasty" case, where f is defined on a Riemannien manifold $\mathcal{M}$ (see Figure 1(a)). Since the support of f is $\mathcal{M}$ instead of $\mathbb{R}^d$, search should be more careful with exploration and exploitation restricted on $\mathcal{M}$. Very often the shape of $\mathcal{M}$ is irregular, which worsens the situation. However, this task arises in many areas [27, 8]. Therefore, it is rather desirable to come up with a general algorithm which can at the best efficiency find good solutions with guaranteed precision.

Simulated annealing (SA) is a classical stochastic optimization method for black-box functions [21]. Experience users usually spend much effort in tuning parameters in SA to accelerate its convergence. Although an adaptive mechanism was introduced by Ingber (1996) [19], yet its applicability is restricted within *random walk metropolis algorithm*. Therefore, its adaptivity is rather limited for complex functions. In present paper, we propose an adaptive stochastic method to handle f on Remannian manifolds. The proposed algorithm is based on *Kernel Adaptive Sequential Monte Carlo* (KASMC) by exploiting the close analogy between SA and sequential Monte Carlo (SMC), and therefore is referred to as KASCM optimizer. Similar to SA, KASMC optimizer increases the determinism of proposal acceptance criterion by monotonically decreasing a temperature parameter. However, in KASMC, sequential tempering is implemented in a SMC instead, which can be considered as a multi-particle simulated annealing [25] with weighting and resampling. At each Markov chain Monte Carlo (MCMC) transition in the SMC, samples are first proposed in a *reproducing kernel Hilbert space* (RKHS) of the embedding space of $\mathcal{M}$, after adaptively tempered acceptance criterion, they are mapped back to the embedding space and $\mathcal{M}$ respectively by two projection steps. The adaptive proposal construction in RKHS is inspired by *MCMC Kameleon* [31]. Meanwhile, we improve MCMC Kameleon with a more effective method for finding pre-images by exploiting local topology consistence between RKHS and the embedding space. The strength of KASMC optimizer stems from its two adaptive tunning mechanisms. The first one is adaptive tempering, which, based on *effective sampling size* (ESS), automatically constructs temperature schedule through cooling; the second one is adaptive MCMC transition in RKHS, which is critical to handle highly nonlinear support of f. Furthermore, based on previous studies, which include the finite-time performance analysis of simulated annealing on continuous domains [22, 23] and convergence property of adaptive SMC [6], we present a finite-time performance guarantee of KASMC optimizer. In our experiments, the proposed KASMC optimizer was evaluated on two practical tasks: rotation optimization on $SO(3)$ manifold for 3D registration and subspace optimization on Stiefel manifold for dimension reduction in classification. According to our empirical results, KASMC optimizer consistently outperforms SA and other non(or weaker)-adaptive counterparts.

Related Work KASMC optimizer can be related to many other work from different perspectives. The following provides a short summary of recent development of relevant study. **Global optimization of black-box functions.** A review of methods for global optimization of black-box functions can be found in [28] or [11]. Out of all work, *Bayesian optimization* [7, 4] and *Hierarchical optimistic optimization* [26] are two research branches which have recently attracted particularly more attention. Bayesian optimization is a *response surface method* with Gaussian process to interpolate the unknown function f. Hierarchical optimistic optimization is an application of *optimism principle* on *Monte Carlo tree search* (MCTS), which partitions search spaces hierarchical and conducts exploration and exploitation at different scales. To the best of our knowledge, no work has been conducted for Bayesian optimization or optimistic optimization on Riemannian manifolds. Only one work on Bayesian optimization with constraints [15] can be, in some cases, extended to Reimannian manifolds. **Linking SA and SMC.** SA and SMC were rarely connected although both of them have been employed for stochastic optimization. Recently a few exceptions emerged [33] and [16], however, most of them are simplified cases of our work. **Adaptive SMC.** Inspired by adaptive MCMC, many adaptive versions of SMC have been also studied. Nevertheless, in SMC, not only is the MCMC transition adaptively tuned, but also the temperature sequence is also adaptively determined [13, 20, 29, 6]. **MCMC on Riemannian Manifold.** A notable work was recently presented by Byrne et al.(2014) [9], who proposed a *Geodesic MCMC* by exploiting geodesics and geometric measure theory. Meanwhile, our method is more general because no explicit geodesic is needed.

This paper makes four contributions. First, we reveal the closeness between SA and SMC, which sheds some light on new understanding on these two techniques. Second, to better fit nonlinear support (here Riemannian manifold), MCMC Kameleon is for the first time extended for adaptive SMC, and we put forward a simpler yet better method for finding pre-images to replace the naive one-step gradient in MCMC Kameleon. Third, a general and fully adaptive stochastic optimization method, KASMC, for black-box functions on Riemannian manifolds is developed, in which no parameter tunning is needed. Last but not least, we derive a rigorous and informative finite-time performance guarantees for the KASMC optimizer.

2 Preliminaries

2.1 Assumptions

Given a function f defined on a Riemannian manifold $\mathcal{M}$, $f : \mathcal{M} \to \mathbb{R}$, we are considering an optimization problem:

$$f^* := \sup_{\theta \in \mathcal{M}} f(\theta) \tag{1}$$

The following assumption will be held throughout the paper:

Assumption 1 *$\mathcal{M}$ is a compact Riemannian manifold embedded in $\mathbb{R}^d$ and it has finite Lebegsgue measure λ. f is a point-wise measurable function on $\mathcal{M}$ and bounded within range* $[0, 1]$.

Obviously, the range restriction in Assumption 1 can be immediately relaxed to any arbitrary bounded function $f'(\theta) \in [\underline{f}, \bar{f}]$ by shifting and rescaling: $f(\theta) = \frac{f'(\theta)-\underline{f}}{\bar{f}-\underline{f}}$.

2.2 Simulated Annealing

SA is a stochastic optimization method and it was originally proposed as an adaption of Metropolis-Hasting (MH) algorithm to obtain samples from thermodynamic systems [21]. Essentially, SA is a time-inhomogeneous MCMC with the stationary distribution modified by a decreasing temperature parameter β. Following Lecchini-Visintini et al. (2007, 2010) [22, 23], we define a density proportional to $[f(\theta) + \delta]^J$:

$$\pi(d\theta; J, \delta) \propto [f(\theta) + \delta]^J \lambda(d\theta) \tag{2}$$

where J is referred to as *inverse temperature* ($J = \frac{1}{\beta}$) and δ is an extra parameter for controlling convergence. Consequently, the acceptance probability of a new sate $\tilde{\theta}$ at the $k + 1$th iteration of SA is:

$$\min \left\{ 1, \frac{q(\theta_k|\tilde{\theta})[f(\tilde{\theta}) + \delta]^{J_{k+1}}}{q(\tilde{\theta}|\theta_k)[f(\theta_k) + \delta]^{J_{k+1}}} \right\} \tag{3}$$

where $q(\cdot|\cdot)$ is a proposal distribution. When use SA, a target J is required ($J_1 < J_2 < \cdot J_k < \cdots J$) and it is usually tuned based on empirical experience. Fortunately, the theory developed by Lecchini-Visintini et al. (2007, 2010) [22, 23] provides a guide of setting J.

Theorem 1 *[23] Let assumption 1 hold. Let $J \geq 1$ and $\delta > 0$, then for any $\alpha \in (0, 1], \epsilon \in (0, 1]$, if*

$$J \geq \frac{1 + \epsilon + \delta}{\epsilon} \left[\log \frac{\sigma}{1 - \sigma} + \log \frac{1}{\alpha} + 2 \log \frac{1 + \delta}{\delta} \right] \tag{4}$$

then

$$P_{\theta_{k(J)}}(\Theta(\epsilon, \alpha)) \geq \sigma - ||P_{\theta_k} - \pi(\cdot; J, \delta)||_{TV} \tag{5}$$

where $\Theta(\epsilon, \alpha)$ is referred to as *the approximate domain optimizer* of f with imprecision ϵ and residual domain α (see following definition):

$$\Theta(\epsilon, \alpha) = \{\theta \in \mathcal{M} | \lambda(\{\theta' \in \Theta | f(\theta' > f(\theta) + \epsilon)\}) \leq \alpha \lambda(\mathcal{M})\} \tag{6}$$

$\theta_{k(J)}$ with distribution $P_{\theta_{k(J)}}$ is the state of the Markov chain at the iteration when the inverse temperature reaches J, $||P_{\theta_{k(J)}} - \pi(\cdot; J, \delta)||_{TV}$ denotes the *total variance distance* between $P_{\theta_{k(J)}}$ and $\pi(\cdot; J, \delta)$ [14]:

$$||P_{\theta_{k(J)}} - \pi(\cdot; J, \delta)||_{TV} = \frac{1}{2} \int |P_{\theta_{k(J)}} - \pi(\cdot; J, \delta)| = \sup_{\theta \in \mathcal{M}} |P_{\theta_{k(J)}}(\theta) - \pi(\theta; J, \delta)| \tag{7}$$

Based on the right hand side of (4), we can also find an optimal δ to minimize the lower bound.

Lecchini-Visintini et al. (2010) [23] studied $||P_{\theta_k} - \pi(\cdot; J, \delta)||_{TV}$ in a very loose way and obtained a bound regardless of $\pi(\cdot; J, \delta)$. The reason of this illusion is that they employed a rather inefficient MH transition (*i.e.* uniform proposal distribution). Indeed, however, the analysis of $||P_{\theta_k} - \pi(\cdot; J, \delta)||_{TV}$ with other proposal distributions in SA context is difficult. In addition, providing only a lower bound of J (4) is not of high practical value since no guide is given for temperature scheduling, which is rather critical to the mixing rate of the Markov chain, and thus $||P_{\theta_k} - \pi(\cdot; J, \delta)||_{TV}$. As a matter of fact, these deficiencies motivate our work.

2.3 SMC and Adaptive SMC

Sequential Monte Carlo (SMC) was original developed in filtering community for state-space models, so it is also referred to as *particle filter* and also its samples are called *particles* [17]. Later, the

Algorithm 1 Sampling using SMC

```
Initialize $p(\mathbf{x}; \boldsymbol{\theta}_0), t \leftarrow 0$
Sample particles $\{\bar{\mathbf{x}}_0^{(s)}\}_{s=1}^S \sim p(\mathbf{x}; \boldsymbol{\theta}_0)$
while ! stop criterion do
    $h \leftarrow 0, \beta_0 \leftarrow 1$
    while $\beta_h < 1$ do
        assign importance weights $\{w^{(s)}\}_{s=1}^S$ to particles according to (??)
        resample particles based on $\{w^{(s)}\}_{s=1}^S$
        compute the step length $\Delta\beta_h$ according to Algorithm ??
        $\beta_{h+1} = \beta_h + \delta\beta$
        $h \leftarrow h + 1$
    end while
    Compute the gradient $\Delta\boldsymbol{\theta}_t$ according to (??)
    $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \eta\Delta\boldsymbol{\theta}_t$
    $t \leftarrow t + 1$
end while
Output:    samples from $\pi_K$
```

usage of SMC was extended as an alternative sampler of MCMC for complex densities or a stochastic optimization method [12, 10, 24]. Instead of sampling from an interval-fixed sate-distribution sequence in filtering, one can construct a synthetic distribution sequence, ordered with increasing complexity or increasing number of data (*e.g.* in Bayesian inference), and use SMC to go through this distribution ladder to sample from a target distribution of interest. SMC sampler was empirically proved to be more successful than MCMC in many cases [10, 24, 32]. A pseudo code of SMC sampler is presented in Algorithm 1.

Here we are going to compare SMC and SA to gain some deeper insights into these two techniques.

Inspired by adaptive MCMC [5],

In Beskos et al.(2013) [6], a *central limit theorem* (CLT) was proved for $\mathbb{E}[\tilde{\pi}_k - \pi_k] = \int(\tilde{\pi}_k(\theta) - \pi_k(\theta))d\theta = \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^N(\tilde{\pi}(\theta^{(n)}) - \pi(\theta^{(n)}))$.

Theorem 2 *[6]: In SMC with adaptive transition scaling and adaptive tempering, at each iteration,*

$$\left\{ N^{\frac{1}{2}}\mathbb{E}[\tilde{\pi}_k - \pi_k] \right\} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{Var}[\pi_k]) \tag{8}$$

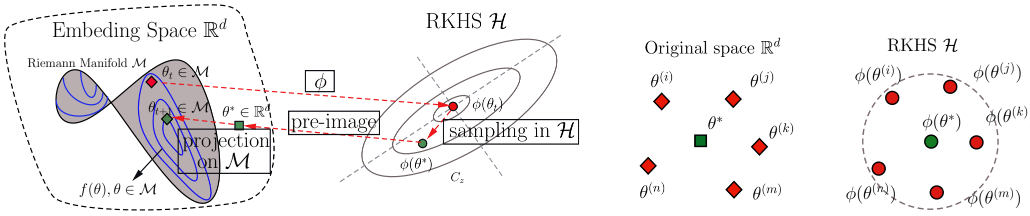
A rule of thumb for CLT is that $N = 30$ is sufficiently large [18]. Therefore, Theorem 2 holds for all cases where the size of the particle set $N \geq 30$. It is suggested by Beskos et al. (2013) [6] that N be chosen of order $\mathcal{O}(d)$.

2.4 Kernel Adaptive MCMC

A kernel adaptive Metropolis Hasting (MH) sampler, called *MCMC Kameleon*, was recently proposed by Sejdinovic et al. (2014) [31]. The key observation of MCMC Kameleon is that the trajectory of the Markov chain is mapped to a *reproducing kernel Hilbert space* (RKHS), where a proposal is constructed based on the feature space covariance of the samples, and then mapped back to the original space.

During its burn-in phase it obtains a subsample $\mathbf{z} = \{z_i\}_{i=1}^n$ of the chain history $\{x_i\}_{i=0}^{t-1}$ at each iteration for updating the proposal distribution $q_{\mathbf{z}}(\cdot \mid x)$ to learn an approximation of the target density π. It does so by applying kernel PCA [30] on $\mathbf{z}$ which results in a low-rank covariance operator $C_{\mathbf{z}}$. Using $\nu^2 C_{\mathbf{z}}$ as a covariance (where ν is a scaling parameter), next a Gaussian measure with mean $k(\cdot, y)$, i.e., $\mathcal{N}(f; k(\cdot, y), \nu^2 C_{\mathbf{z}})$, is defined in "density form". Samples f from this measure are subsequently used to obtain target proposals x^*.

The difficulty of Kameleon MCMC in generating target samples is that f cannot be mapped back by a closed-form solution to its pre-image in $\mathcal{X}$. Yet, it is assumed that there exists a point $x^* \in \mathcal{X}$



(a)

(b)

Figure 1: An overview of kernel Adaptive MCMC transition and two inverse projection steps.

whose canonical feature map φ is close to f under the RKHS norm. By formulating this as a non-convex optimization problem

$$\underset{x \in \mathcal{X}}{\operatorname{argmin}} \|k(\cdot, x) - f\|_{\mathcal{H}}^2 \tag{9}$$

and taking a gradient step along the cost function

$$g(x) = k(x,x) - 2k(x,y) - 2\sum_{i=1}^{n} \beta_i \left[k(x, z_i) - \mu_{\mathbf{z}}(x)\right] \tag{10}$$

a new target proposal x^* is given by

$$x^* = y - \eta \nabla_x g(x)|_{x=y} + \xi \tag{11}$$

where β is a vector of coefficients, η is the gradient step size, and $\xi \sim \mathcal{N}(0, \gamma^2 I)$ an additional isotropic exploration term after the gradient. The complete Kameleon MCMC algorithm [**?**] then is

- at each iteration $t + 1$
  1. obtain a subsample $\mathbf{z} = \{z_i\}_{i=1}^n$ of the chain history $\{x_i\}_{i=0}^{t-1}$,
  2. sample $x^* \sim q_{\mathbf{z}}(\cdot \mid x_t) = \mathcal{N}(x_t, \gamma^2 I + \nu^2 M_{\mathbf{z},x_t} H M_{\mathbf{z},x_t}^T)$,
  3. accept x^* with MH acceptance probability $\alpha(x,y) = min\left\{1, \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}\right\}$.

where $M_{\mathbf{z},y} = 2\eta \left[\nabla_x k(x, z_1)|_{x=y}, \ldots, \nabla_x k(x, z_n)|_{x=y}\right])]$ is the kernel gradient matrix which is obtained from the gradient of (10) at y, γ a noise parameter, and H an $n \times n$ centering matrix.

As of its Gaussian proposals (both, x^* and f) Kameleon MCMC calculates closed form solutions and is analytically tractable.

3 KASMC Optimizer

3.1 An Overview

Adaptive MCMC transition in RKHS

Adaptive Tempering

3.2 Finding Pre-images in the Embedding Space

Mapping $\phi(\theta^*)$ back to θ^* is referred to as *pre-image* problem and formulated as:

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^d} \|\phi(\theta) - \phi(\theta^*)\|_{\mathcal{H}}^2 \tag{12}$$

In general, (12) is a non-convex minimization problem and therefore difficult to solve. Sejdinovic et al. (2014) [31] employed a simple yet rather unreliable short-cut to get an approximate solution: by moving a single descent step along the gradient of (12). Although by doing this they can obtain a simple multivariate Gaussian in $\mathbb{R}^d$ (**??**), yet it can only be applied on differentiable kernels. Here, instead, we put forward a new and more general method for pre-images by exploiting the *topology consistence* within a small local area between the original space and the Hilbert space. Basically, we assume that within a small area the spatial configuration of points will not change too much after being projected into a Hilbert space (Figure 1(b)). The method is quite simple as follows:

$$\theta^* = \sum_{u=1:\phi(z^{(u)})\in\mathbf{N}(\phi(\theta^*))}^{U} w^{(u)}z^{(u)} \Bigg/ \sum_{u=1:\phi(z^{(u)})\in\mathbf{N}(\phi(\theta^*))}^{U} w^{(u)} \tag{13}$$

where $\mathbf{N}(\phi(\theta^*))$ denotes the neighbourhood of $\phi(\theta^*)$ in $\mathcal{H}$ and it was obtained by ranking values of $w^{(u)} = K(z^{(u)}, \theta^*)$ and taking the top U ones.

3.3 Projection onto $\mathcal{M}$

After begin projected to the embedding space, θ^* usually still lays outside of $\mathcal{M}$. Therefore, we need one extra projection $\mathbf{P}$ to map θ^* onto $\mathcal{M}$:

$$\mathbf{P}(\theta^*) = \arg\min_{\theta\in\mathcal{M}} \|\theta - \theta^*\|^2 \tag{14}$$

In general, to solve the above convex objective function, we can resort to numerical optimization methods on Riemannnian manifolds [2]. In our experiments, $SO(3)$ manifold and Stiefel manifold are considered, projection on them actually can be more easily computed.

Theorem 3 *Given an arbitrary matrix $A \in \mathbb{R}^{p\times r}$, its projection onto the Stiefel manifold $St(r,p)$ is:*

$$\mathbf{P}_{St}(A) = UV^\top, \quad \textit{where U and V are left and right eigenvectors of B respectively.} \tag{15}$$

Theorem 4 *Given an arbitrary matrix $B \in \mathbb{R}^{p\times p}$, its projection onto $SO(3)$ manifold is:*

$$\mathbf{P}_{SO}(B) = UV^\top, \quad \textit{where U and V are left and right eigenvectors of A respectively.} \tag{16}$$

Proof See [3].

3.4 A Finite-time Performance Guarantee

Corollary 1 *Let Assumption 1 and the conditions in Theorem 1 hold, when J satisfies the lower bound defined in (4),*

$$P_{\theta_{k(J)}}(\Theta(\epsilon,\alpha)) \geq \left[\sigma - \eta(N^{\frac{3}{4}} + (N+1)^{\frac{3}{4}})\mathbf{Std}[\pi(\cdot; J, \delta)]\right] \cdot \text{erf}(\frac{\eta}{\sqrt{2}})^2 \tag{17}$$

where $\eta \in \mathbb{R}^+$, N is the size of the particle set, erf$(\cdot)$ denotes *Gauss error function* and $\mathbf{Std}[\tau]$ is the standard deviation of a distribution τ.

Proof See Appendix

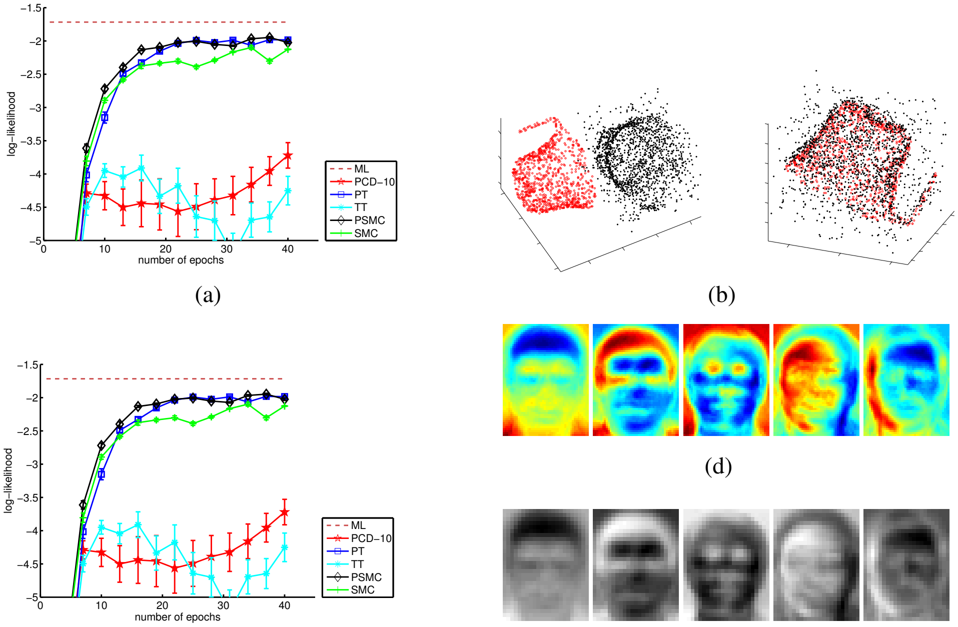
We can see (17) provides a more informative bound than (5). In particular, (17) explicitly invovles $\mathbf{Std}[\pi(\cdot; J, \delta)]$, which is an important factor to analysis. For instance, if f is a unimodal function and J is sufficiently large, $\mathbf{Std}[\pi(\cdot; J, \delta)]$ will be quite close to 0.

Corollary 2 *Let Assumption 1 and the conditions in Theorem 1 hold, when J satisfies the lower bound defined in (4),*

$$P_{\theta_{k(J)}}(\Theta(\epsilon,\alpha) \in \mathcal{S}) \geq 1 - \left\{1 - \left[\sigma - \eta(N^{\frac{3}{4}} + (N+1)^{\frac{3}{4}})\mathbf{Std}[\pi(\cdot; J, \delta)]\right] \cdot \text{erf}(\frac{\eta}{\sqrt{2}})^2\right\}^N \tag{18}$$

Proof Since particles in KASMC are not independent, instead,

$$P_{\theta_{k(J)}}(\Theta(\epsilon,\alpha) \in \mathcal{S}) = 1 - P(\forall\theta^{(n)} \in \mathcal{S}, \theta^{(n)} \neq \Theta(\epsilon,\alpha)) \geq 1 - [1 - P_{\theta_{k(J)}}(\Theta(\epsilon,\alpha)]^N \tag{19}$$



(a) (b) (d) (c) (e)

Figure 2: (a) (c) Performance comparison of five optimization algorithms in the rotation optimization experiment and subspace optimization experiment respectively. (b) Optimal rotation found by KASMC for 3D registration. (d) 5 bases of the optimal subspace. (e) 5 bases computed using PCA.

4 Experiments

Two experiments were conducted for practical tasks: finding optimal rotation and finding optimal subspace. For comparison, besides KASMC, we also tested other four algorithms: *i*.N independent simulated annealing (N-SA); *ii*.SMC; *iii*.Adaptive SMC (ASMC); *iv*.KASMC with one gradient move for computing pre-images (KASMC-P). N in N-SA is set to be equivalent to the number of particles in SMCs. Since there is no adaptive tempering in N-SA and SMC, their temperature sequences are set by using popular logarithmic schedule $J_k = \lfloor \log k \rfloor + 1$ until it hits the target J. Except pre-image computation step, other components in KASMC-P and KASMC are set up in the same way. In addition, the same projection procedure onto manifolds are used in all five algorithms.

4.1 Rotation Optimization on $SO(3)$ Manifold

The first experiment is for 3D point cloud registration, which is difficult without prior knowledge of correspondence. Two point clouds X and Y were generated by first sampling a point set from a 3D mesh file and then applying two different 3D transformations on it. Since there is no noisy, translation can be ignored by simply shifting the centers of two point clouds to the origin. Then the objective function on only rotation can be defined as:

$$f(\theta) = \frac{1}{1 + \mathcal{L}(\theta)} \qquad \mathcal{L}(\theta) = \sum_{m=1}^{M} (x^{(c(m))} - \theta y^{(m)}), \qquad \theta \in SO(3) \tag{20}$$

where $M = |Y|$, $\{x^{(c(m))}, y^{(m)}\}$ is a pair of matching points in X and Y respectively, and the matching of $\theta y^{(m)}$ is found by searching a point in X, $x^{(c(m))}$, which is closest to it. Obviously, f is a black-box function on $SO(3)$ and bounded within the range $(0, 1]$.

4.2 Subspace Optimization on Stiefel Manifold

The second experiment is for wearing-glass classification (*i.e.* determine whether a face image has glass), and the target is to find optimal subspace of human faces for minimizing misclassification loss. For simplicity, we used *K-nearest-neighbour* (KNN) for classification. We used images from the *ORL face database* [1] and shrank them to $\frac{1}{4}$ of the original size (*i.e.* 28×23). Half of them were used for training and the other half for testing. Here our interest is a 5 dimensional subspace, therefore, the search space is a Stiefel manifold $St(5, d = 28 \times 23)$.

$$f(\theta) = \frac{1}{1 + \mathcal{L}(\theta)} \qquad \mathcal{L}(\theta) = \text{misclassification error}, \qquad \theta \in St(5, d) \tag{21}$$

Obviously, f is a black-box function on $St(5, d)$. Assume there exist M test instances then $f(\theta) \in [\frac{1}{M+1}, 1]$. Furthermore, we visualize 5 discriminative bases in Figure 2(d). We can see that all 5 bases focus around eyes, which differ very much from the ones from using PCA (Figure 2(e)).

5 Conclusion

A new stochastic optimization method was presented for black-box functions on Riemannian manifolds. Many state-of-the-art studies were exploited. The proposed KASMC optimizer outperforms classic methods in both effectiveness and efficiency. As a possible future work direction, we are going to investigate how KASMC can be combined other strategies: *e.g.* Bayesian optimization or optimistic optimization.

Appendix

Based on Theorem 2, given a particle set $\mathcal{S} = \{\theta^{(n)}\}_{n=1}^{N}, N \geq 30$, when the inverse temperature reaches J,

$$\forall \mathcal{S} \subseteq \mathcal{M}, \qquad \frac{1}{N} \sum_{n=1:\theta^{(n)} \in \mathcal{M}}^{N} (P_{\theta_{k(J)}}(\theta^{(n)}) - \pi(\theta^{(n)}; J, \delta)) \sim \mathcal{N}(0, \Sigma) \tag{22}$$

where $\Sigma = N^{-\frac{1}{2}} \mathbf{Var}[\pi(\cdot; J, \delta)]$. By using *cumulative density function* (CDF) , we have $P(E) = \text{erf}\left(\frac{\eta}{\sqrt{2}}\right)$, where E denotes the following event:

$$-\eta N^{-\frac{1}{4}} \mathbf{Std}[\pi(\cdot; J, \delta)] \leq \frac{1}{N} \sum_{n=1:\theta^{(n)} \in \mathcal{M}}^{N} (P_{\theta_{k(J)}}(\theta^{(n)}) - \pi(\theta^{(n)}; J, \delta)) \leq \eta N^{-\frac{1}{4}} \mathbf{Std}[\pi(\cdot; J, \delta)] \tag{23}$$

$$\iff -\eta N^{\frac{3}{4}} \mathbf{Std}[\pi(\cdot; J, \delta)] \leq \sum_{n=1:\theta^{(n)} \in \mathcal{M}}^{N} (P_{\theta_{k(J)}}(\theta^{(n)}) - \pi(\theta^{(n)}; J, \delta)) \leq \eta N^{\frac{3}{4}} \mathbf{Std}[\pi(\cdot; J, \delta)] \tag{24}$$

With the same probability ($P(E^\dagger) = \text{erf}\left(\frac{\eta}{\sqrt{2}}\right)$), (24) also holds for $\mathcal{S}^\dagger = \mathcal{S} \cup \theta^\dagger$, where $\theta^\dagger \in \mathcal{M} - \mathcal{S}$:

$$-\eta(N+1)^{\frac{3}{4}} \mathbf{Std}[\pi(\cdot; J, \delta)] \leq \sum_{n=1:\theta^{(n)} \in \mathcal{S} \cup \theta^\dagger}^{N+1} (P_{\theta_{k(J)}}(\theta^{(n)}) - \pi(\theta^{(n)}; J, \delta)) \leq \eta(N+1)^{\frac{3}{4}} \mathbf{Std}[\pi(\cdot; J, \delta)] \tag{25}$$

therefore, with probability $P(E^*) = \text{erf}(\frac{\eta}{\sqrt{2}})^2$, where E^* denotes the following event:

$$\left| P_{\theta_{k(J)}}(\theta^\dagger) - \pi(\theta^\dagger; J, \delta) \right| \leq \eta(N^{\frac{3}{4}} + (N+1)^{\frac{3}{4}}) \mathbf{Std}[\pi(\cdot; J, \delta)] \tag{26}$$

Since $\mathcal{S}$ and therefore $\theta^\dagger$ are arbitrary, E^* is equivalent to

$$\sup_{\theta \in \mathcal{M}} \left| P_{\theta_{k(J)}}(\theta) - \pi(\theta; J, \delta) \right| \leq \eta(N^{\frac{3}{4}} + (N+1)^{\frac{3}{4}}) \mathbf{Std}[\pi(\cdot; J, \delta)] \tag{27}$$

Therefore, when J satisfies the lower bound defined in (4),

$$P_{\theta_{k(J)}}(\Theta(\epsilon, \alpha)) \geq \left[\sigma - \eta(N^{\frac{3}{4}} + (N+1)^{\frac{3}{4}}) \mathbf{Std}[\pi(\cdot; J, \delta)] \right] \cdot \text{erf}(\frac{\eta}{\sqrt{2}})^2 \tag{28}$$

Acknowledgement

The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 (Specific Programme Cooperation, Theme 3, Information and Communication Technologies) under grant agreement no. 270273, Xperience.

References

[1] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.

[2] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

[3] P.-A. Absil and J. Malick. Projection-like retractions on matrix manifolds. *SIAM J. on Optimization*, 22(1):135–158, Jan. 2012.

[4] R. P. Adams, Z. Ghahramani, M. W. Hoffman, J. Snoek, and K. Swersky. NIPS workshop on Bayesian Optimization in Academia and Industries, 2014.

[5] C. Andrieu and J. Thoms. A tutorial on adaptive mcmc. *Statistics and Computing*, 18(4):343–373, 2008.

[6] A. Beskos, A. Jasra, N. Kantas, and A. Thiery. On the convergence of adaptive sequential monte carlo methods. Technical Report arXiv:1306.6462v3, arXiv.org, 2013.

[7] E. Brochu, V. M. Cora, and N. de Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. Technical Report arXiv:1012.2599, arXiv.org, December 2010.

[8] S. Byrne and M. Girolami. Geodesic monte carlo on embedded manifolds. *Scandinavian Journal of Statistics*, 40(4):825–845, 2013.

[9] S. Byrne and M. Girolami. Geodesic monte carlo on embedded manifolds. *Scandinavian Journal of Statistics*, 40(4):825–845, 2013.

[10] N. Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3):539–552, 2002.

[11] A. R. Conn, K. Scheinberg, and L. N. Vicente. *Introduction to Derivative-Free Optimization*. Society for Industrial and Applied Mathematics, 2009.

[12] P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, June 2006.

[13] P. Del Moral, A. Doucet, and A. Jasra. An adaptive sequential monte carlo method for approximate bayesian computation. *Statistics and Computing*, 22(5):1009–1020, 2012.

[14] L. Devroye and L. Györfi. *Nonparametric density estimation: the L1 view*. Wiley series in probability and mathematical statistics. 1985.

[15] J. Gardner, M. Kusner, K. Q. Weinberger, J. Cunningham, and Z. Xu. Bayesian optimization with inequality constraints. In *ICML*, 2014.

[16] J. Geweke and B. Frischknecht. Exact optimization by means of sequentially adaptive Bayesian learning, 2014. unpulished.

[17] N. Gordon, D. Salmond, and A. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEEE Proceedings F, Radar and Signal Processing*, 140(2):107–113, 1993.

[18] C. M. Grinstead and L. J. Snell. *Grinstead and Snell's Introduction to Probability*. American Mathematical Society, 2006.

[19] L. Ingber. Adaptive simulated annealing (asa): Lessons learned. *Control and Cybernetics*, 25:33–54, 1996.

[20] A. Jasra, D. A. Stephens, A. Doucet, and T. Tsagaris. Inference for lévy-driven stochastic volatility models via adaptive sequential monte carlo. *Scandinavian Journal of Statistics*, 38(1):1–22, 2011.

[21] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *SCIENCE*, 220(4598):671–680, 1983.

[22] A. Lecchini-visintini, J. Lygeros, and J. Maciejowski. Simulated annealing: Rigorous finite-time guarantees for optimization on continuous domains. In *Advances in Neural Information Processing Systems (NIPS)*, pages 865–872. MIT Press, 2007.

[23] A. Lecchini-Visintini, J. Lygeros, and J. M. Maciejowski. Stochastic optimization on continuous domains with finite-time guarantees by markov chain monte carlo methods. *Automatic Control, IEEE Transactions on*, 55(12):2858–2863, Dec 2010.

[24] J. Míguez, D. Crisan, and P. Djurić. On the convergence of two sequential monte carlo methods for maximum a posteriori sequence estimation and stochastic global optimization. *Statistics and Computing*, 23(1):91–107, 2013.

[25] O. Molvalioglu, Z. Zabinsky, and W. Kohn. Multi-particle simulated annealing. In A. Trn and J. ilinskas, editors, *Models and Algorithms for Global Optimization*, volume 4 of *Optimization and Its Applications*, pages 215–222. Springer US, 2007.

[26] R. Munos. From bandits to monte-carlo tree search: The optimistic principle applied to optimization and planning. *Foundations and Trends in Machine Learning*, 7(1):1–129, 2014.

[27] C. Papazov and D. Burschka. Stochastic global optimization for robust point set registration. *Comput. Vis. Image Underst.*, 115(12):1598–1609, Dec. 2011.

[28] L. Rios and N. Sahinidis. Derivative-free optimization: a review of algorithms and comparison of software implementations. *Journal of Global Optimization*, 56(3):1247–1293, 2013.

[29] C. Schäfer and N. Chopin. Sequential monte carlo on large binary sampling spaces. *Statistics and Computing*, 23(2):163–184, 2013.

[30] B. Schölkopf, A. Smola, and K.-R. Müller. Kernel Principal Component Analysis. In *Artificial Neural Networks — ICANN'97*, pages 583–588. Springer, 1997.

[31] D. Sejdinovic, H. Strathmann, M. L. Garcia, C. Andrieu, and A. Gretton. Kernel Adaptive Metropolis-Hastings. In *ICML*, pages 1665–1673, 2014.

[32] H. Xiong, S. Szedmak, and J. Piater. Towards Maximum Likelihood: Learning Undirected Graphical Models using Persistent Sequential Monte Carlo. In *6th Asian Conference on Machine Learning*, 2014.

[33] E. Zhou and X. Chen. Sequential monte carlo simulated annealing. *Journal of Global Optimization*, 55(1):101–124, 2013.