POLITECNICO DI MILANO

DEPARTMENT OF ELECTRONICS AND INFORMATION
http://www.dei.polimi.it

AIRLab Artificial Intelligence and Robotics Laboratory
http://www.airlab.elet.polimi.it

Graphical Models and Model-Based Search Algorithms

Luigi Malagò
Politecnico di Milano → Università degli Studi di Milano

Workshop on Graphical Models,
The Fields Institute, Toronto, 18 April 2012

Aim of the Talk

1. Present an application of graphical (log-linear) models in model-based meta-heuristics for optimization

Aim of the Talk

1. Present an application of graphical (log-linear) models in model-based meta-heuristics for optimization
2. Discuss new approaches to optimization and model selectionbased on natural gradient and linear regression

Aim of the Talk

1. Present an application of graphical (log-linear) models in model-based meta-heuristics for optimization
2. Discuss new approaches to optimization and model selectionbased on natural gradient and linear regression

Outline

- Brief introduction to Model-Based Search (MBS)

Aim of the Talk

1. Present an application of graphical (log-linear) models in model-based meta-heuristics for optimization
2. Discuss new approaches to optimization and model selectionbased on natural gradient and linear regression

Outline

- Brief introduction to Model-Based Search (MBS)
- Why graphical models in such context?

Aim of the Talk

1. Present an application of graphical (log-linear) models in model-based meta-heuristics for optimization
2. Discuss new approaches to optimization and model selectionbased on natural gradient and linear regression

Outline

- Brief introduction to Model-Based Search (MBS)
- Why graphical models in such context?
- Main issues and open problems

Aim of the Talk

1. Present an application of graphical (log-linear) models in model-based meta-heuristics for optimization
2. Discuss new approaches to optimization and model selectionbased on natural gradient and linear regression

Outline

- Brief introduction to Model-Based Search (MBS)
- Why graphical models in such context?
- Main issues and open problems
- Fitness modelling and natural gradient

Aim of the Talk

1. Present an application of graphical (log-linear) models in model-based meta-heuristics for optimization
2. Discuss new approaches to optimization and model selectionbased on natural gradient and linear regression

Outline

- Brief introduction to Model-Based Search (MBS)
- Why graphical models in such context?
- Main issues and open problems
- Fitness modelling and natural gradient
- Model selection and linear regression

Model-Based Optimization

- Model-based Search (MBS) (Zlochin et al., 2004) is paradigm in optimization based on the idea of finding the minimum by identifying a proper sequence of densities in a statistical model

Model-Based Optimization

- Model-based Search (MBS) (Zlochin et al., 2004) is paradigm in optimization based on the idea of finding the minimum by identifying a proper sequence of densities in a statistical model
- Black-box context: the analytic formula of the function to be optimized may be unknown

Model-Based Optimization

- Model-based Search (MBS) (Zlochin et al., 2004) is paradigm in optimization based on the idea of finding the minimum by identifying a proper sequence of densities in a statistical model
- Black-box context: the analytic formula of the function to be optimized may be unknown
- Some examples of MBS (and related techniques)
  - Evolutionary computation: EDAs (Larrañaga and Lozano, 2002), GAs (Holland, 1975), ACO (Dorigo, 1992), ESs (Rechenberg, 1960), etc.
  - Gradient descent: CMA-ES (Hansen and Ostermeier, 2001), NES (Wierstra et al., 2008), SGD (Robbins and Monro, 1951)
  - Boltzmann distribution and Gibbs sampler (Geman and Geman, 1984)
  - Simulated Annealing and Boltzmann Machines (Aarts and Korst, 1989)
  - The Cross-Entropy method (Rubinstein, 1997)
  - *LP relaxation in pseudo-Boolean optimization (Boros and Hammer, 2001)*

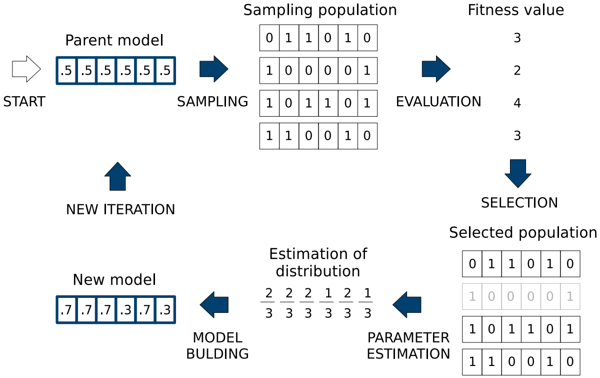
An Example of EDA: UMDA and OneMax

OneMax	Feasible solution	$x = (x_1, \ldots, x_n),\ x_i \in \{0, 1\}$
	Function to maximize	$f(x) = \sum_{i=1}^{n} x_i$
	Statistical model	$p(x) = \Pi_{i=1}^{n} p_i(x_i)$

An Example of EDA: UMDA and OneMax

OneMax

OneMax	Feasible solution	$x = (x_1, \ldots, x_n),\ x_i \in \{0, 1\}$
	Function to maximize	$f(x) = \sum_{i=1}^{n} x_i$
	Statistical model	$p(x) = \Pi_{i=1}^{n} p_i(x_i)$



Estimation of Distribution Algorithms

Let $\mathcal{P}$ be a sample (multiset) of candidate solutions to the optimization problem, and let p a probability distribution

The basic iteration of and EDA consists of

$$\mathcal{P}^t \xrightarrow{\text{selection}} \mathcal{P}^t_s \xrightarrow{\text{estimation}} p^t \xrightarrow{\text{sampling}} \mathcal{P}^{t+1}$$

Estimation of Distribution Algorithms

Let $\mathcal{P}$ be a sample (multiset) of candidate solutions to the optimization problem, and let p a probability distribution

The basic iteration of and EDA consists of

$$\mathcal{P}^t \xrightarrow{\text{selection}} \mathcal{P}_s^t \xrightarrow{\text{estimation}} p^t \xrightarrow{\text{sampling}} \mathcal{P}^{t+1}$$

If we rearrange the elements, we get

$$p^t \xrightarrow{\text{sampling}} \mathcal{P}^{t+1} \xrightarrow{\text{selection}} \mathcal{P}_s^{t+1} \xrightarrow{\text{estimation}} p^{t+1} \qquad p^t \in \mathcal{M}$$

Estimation of Distribution Algorithms

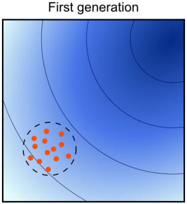
Let $\mathcal{P}$ be a sample (multiset) of candidate solutions to the optimization problem, and let p a probability distribution

The basic iteration of and EDA consists of

$$\mathcal{P}^t \xrightarrow{\text{selection}} \mathcal{P}_s^t \xrightarrow{\text{estimation}} p^t \xrightarrow{\text{sampling}} \mathcal{P}^{t+1}$$

If we rearrange the elements, we get

$$p^t \xrightarrow{\text{sampling}} \mathcal{P}^{t+1} \xrightarrow{\text{selection}} \mathcal{P}_s^{t+1} \xrightarrow{\text{estimation}} p^{t+1} \qquad p^t \in \mathcal{M}$$



Estimation of Distribution Algorithms

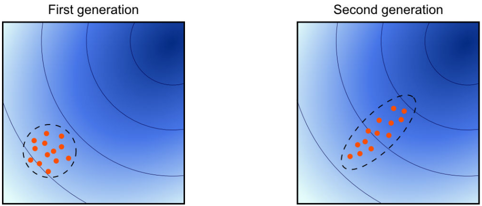
Let $\mathcal{P}$ be a sample (multiset) of candidate solutions to the optimization problem, and let p a probability distribution

The basic iteration of and EDA consists of

$$\mathcal{P}^t \xrightarrow{\text{selection}} \mathcal{P}_s^t \xrightarrow{\text{estimation}} p^t \xrightarrow{\text{sampling}} \mathcal{P}^{t+1}$$

If we rearrange the elements, we get

$$p^t \xrightarrow{\text{sampling}} \mathcal{P}^{t+1} \xrightarrow{\text{selection}} \mathcal{P}_s^{t+1} \xrightarrow{\text{estimation}} p^{t+1} \qquad p^t \in \mathcal{M}$$



Estimation of Distribution Algorithms

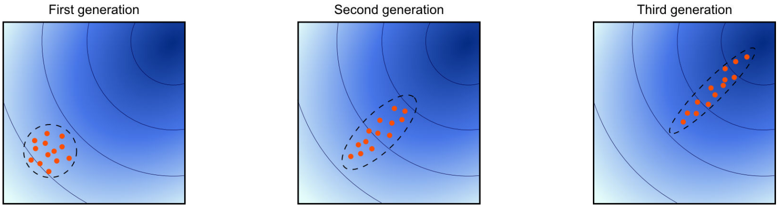
Let $\mathcal{P}$ be a sample (multiset) of candidate solutions to the optimization problem, and let p a probability distribution

The basic iteration of and EDA consists of

$$\mathcal{P}^t \xrightarrow{\text{selection}} \mathcal{P}_s^t \xrightarrow{\text{estimation}} p^t \xrightarrow{\text{sampling}} \mathcal{P}^{t+1}$$

If we rearrange the elements, we get

$$p^t \xrightarrow{\text{sampling}} \mathcal{P}^{t+1} \xrightarrow{\text{selection}} \mathcal{P}_s^{t+1} \xrightarrow{\text{estimation}} p^{t+1} \qquad p^t \in \mathcal{M}$$



A Unifying Perspective for MBS

- Given then original optimization problem $\min_{x \in \Omega} \ f(x)$, we introduce the minimization of the stochastic relaxation $\min_{p \in \mathcal{M}} \ \mathbb{E}_p[f]$ (M., Matteucci and Pistone, 2011)
- We move the search to the space of probability distribution $\mathcal{S}$

A Unifying Perspective for MBS

- Given then original optimization problem $\min_{x \in \Omega} f(x)$, we introduce the minimization of the stochastic relaxation $\min_{p \in \mathcal{M}} \mathbb{E}_p[f]$ (M., Matteucci and Pistone, 2011)
- We move the search to the space of probability distribution $\mathcal{S}$
- Candidate optimal solutions for the original problem can be obtained by sampling the solution of the relaxed problem
- Under proper choice of $\mathcal{M} \subset \mathcal{S}$ the two problems are equivalent

A Unifying Perspective for MBS

- Given then original optimization problem $\min_{x \in \Omega} f(x)$, we introduce the minimization of the stochastic relaxation $\min_{p \in \mathcal{M}} \mathbb{E}_p[f]$ (M., Matteucci and Pistone, 2011)
- We move the search to the space of probability distribution $\mathcal{S}$
- Candidate optimal solutions for the original problem can be obtained by sampling the solution of the relaxed problem
- Under proper choice of $\mathcal{M} \subset \mathcal{S}$ the two problems are equivalent

The relaxed problem can be solved in different ways, e.g, by

- Estimation of distribution (EDAs: Larrañaga and Lozano, 2002)
- Gradient descent (NES: Wierstra et al., 2008)
- Fitness modelling (DEUM framework: Shakya et al., 2005)

Checklist for Model-based Algorithms

- a family of statistical model
- a model selection algorithm
- an estimation algorithm
- a sampling algorithm

Checklist for Model-based Algorithms

- a family of statistical model → Bayesian Networks
- a model selection algorithm → Search+score (BIC/MDL)
- an estimation algorithm → Estimate conditional prob.
- a sampling algorithm → Direct sampling

Checklist for Model-based Algorithms

- a family of statistical model → Bayesian Networks
- a model selection algorithm → Search+score (BIC/MDL)
- an estimation algorithm → Estimate conditional prob.
- a sampling algorithm → Direct sampling

Almost all model-based algorithms employ graphical models, since they provide nice factorizations for the joint probability distribution

EDAs for Discrete Optimization

- Independence model: UMDA (Mühlenbein and Paaß, 1996), PBIL (Baluja, 1994), cGA (Harik, Lobo and Goldberg, 1997)
- Chain: MIMIC (De Bonet, Isbell and Viola, 1997)
- Trees: COMIT (Baluja and Davies (1997)
- Forests: BMDA (Pelikan and Mühlenbein, 1999)
- Clusters of variables: ECGA (Harik, 1999)
- Bayesian Networks: BOA (Pelikan, Goldberg and Cantú-Paz, 2000), EBNA (Etxeberria and Larranãga, 1999), LFDA (Mühlenbein and Mahnig, 1999), hBOA (Pelikan, 2005)
- Markov Random Fields: MN-EDA (Santana, 2005), MOA (Shakya and Santana, 2008)

For a review, see Hauschild and Pelikan (2011)

Directed vs Undirected Graphical Models

Bayesian Networks

- – Learning is hard
- + Estimation is easy
- + Sampling is easy

Directed vs Undirected Graphical Models

Bayesian Networks	Markov Random Fields
– Learning is hard	– Learning is hard
+ Estimation is easy	~ Estimation is not trivial
+ Sampling is easy	~ Sampling is not trivial

State of the art EDAs employ BNs together with decision trees hBOA (Pelikan, 2005)

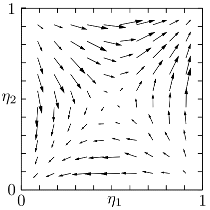
We are interested in MRFs (log-linear models)

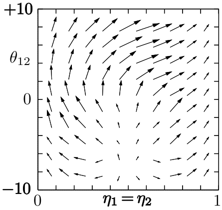
Open Issues

- The choice of $\mathcal{M}$ is crucial in MBS
- Critical points for $\mathbb{E}_p[f]$ imply convergence to local minima

Open Issues

- The choice of $\mathcal{M}$ is crucial in MBS
- Critical points for $\mathbb{E}_p[f]$ imply convergence to local minima





- Efficient methods in the high-dimensional setting (number of variables   100-1K)

The Exponential Family

- We choose models from the exponential family $\mathcal{E}$

$$p(x;\theta) = \exp\left(\sum_{i=1}^{k} \theta_i T_i(x) - \psi(\theta)\right)$$

  - sufficient statistics $T_1(x), \dots, T_k(x)$
  - natural parameters $\theta = (\theta_1, \dots, \theta_k) \in \Theta$
  - log-partition function $\psi(\theta)$

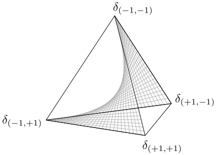
The Exponential Family

- We choose models from the exponential family $\mathcal{E}$

$$p(x;\theta) = \exp\left(\sum_{i=1}^{k}\theta_i T_i(x) - \psi(\theta)\right)$$

  - sufficient statistics $T_1(x), \ldots, T_k(x)$
  - natural parameters $\theta = (\theta_1, \ldots, \theta_k) \in \Theta$
  - log-partition function $\psi(\theta)$

Two parameterizations play a fundamental role (Amari, 2001)



Raw parameters

$$\rho = (\mathbb{P}(X = x))$$

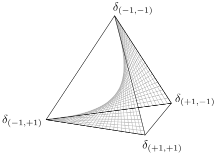
The Exponential Family

- We choose models from the exponential family $\mathcal{E}$

$$p(x;\theta) = \exp\left(\sum_{i=1}^{k}\theta_i T_i(x) - \psi(\theta)\right)$$

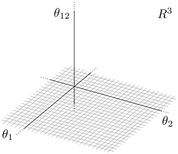
  - sufficient statistics $T_1(x), \ldots, T_k(x)$
  - natural parameters $\theta = (\theta_1, \ldots, \theta_k) \in \Theta$
  - log-partition function $\psi(\theta)$

Two parameterizations play a fundamental role (Amari, 2001)



Raw parameters

$\rho = (\mathbb{P}(X = x))$



Natural parameters

$\theta \in \Theta$

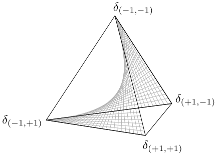
The Exponential Family

- We choose models from the exponential family $\mathcal{E}$

$$p(x;\theta) = \exp\left(\sum_{i=1}^{k} \theta_i T_i(x) - \psi(\theta)\right)$$

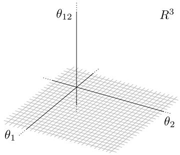
  - sufficient statistics $T_1(x), \dots, T_k(x)$
  - natural parameters $\theta = (\theta_1, \dots, \theta_k) \in \Theta$
  - log-partition function $\psi(\theta)$

Two parameterizations play a fundamental role (Amari, 2001)



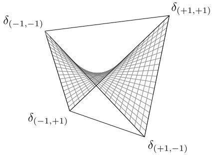
Raw parameters

$\rho = (\mathbb{P}(X = x))$



Natural parameters

$\theta \in \Theta$



Expectation parameters

$\eta = \nabla\psi(\theta) = \mathbb{E}_\theta[T(x)]$

The Gibbs Distribution (Hwang, 1980; Geman and Geman, 1984)

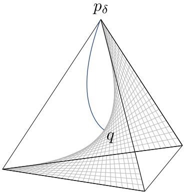
- Given q, the curve following $\nabla \mathbb{E}_p[f]$ is an exponential family

$$p(x;\theta) = \frac{qe^{-\beta f}}{\mathbb{E}_q[e^{-\beta f}]}, \quad \beta > 0$$

- The set of distributions is not weakly closed

$$\lim_{\beta \to 0} p(x;\beta) = q$$

$$\lim_{\beta \to \infty} p(x;\beta) = p_\delta$$



The Gibbs Distribution
(Hwang, 1980; Geman and Geman, 1984)

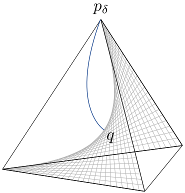
- Given q, the curve following $\nabla \mathbb{E}_p[f]$ is an exponential family

$$p(x;\theta) = \frac{qe^{-\beta f}}{\mathbb{E}_q[e^{-\beta f}]}, \quad \beta > 0$$

- The set of distributions is not weakly closed

$$\lim_{\beta \to 0} p(x;\beta) = q$$

$$\lim_{\beta \to \infty} p(x;\beta) = p_\delta$$



- Since $\nabla \mathrm{E}_\beta[f] = -\mathrm{Var}_\beta(f) < 0$, the expected value of f decreases monotonically

The Gibbs Distribution (Hwang, 1980; Geman and Geman, 1984)

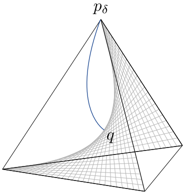
- Given q, the curve following $\nabla \mathbb{E}_p[f]$ is an exponential family

$$p(x;\theta) = \frac{qe^{-\beta f}}{\mathbb{E}_q[e^{-\beta f}]}, \quad \beta > 0$$

- The set of distributions is not weakly closed

$$\lim_{\beta \to 0} p(x;\beta) = q$$

$$\lim_{\beta \to \infty} p(x;\beta) = p_\delta$$



- Since $\nabla \mathrm{E}_\beta[f] = -\mathrm{Var}_\beta(f) < 0$, the expected value of f decreases monotonically

Evaluating the partition function is computationally unfeasible

Geometry of the Exponential Family

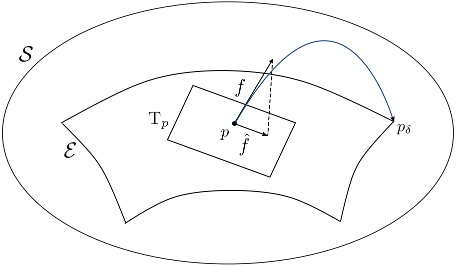
- A statistical model can be modeled as a manifold of distributions by introducing an affine chart in p
- The tangent space in p is defined by $\mathrm{T}_p = \{v : \mathbb{E}_p[v] = 0\}$

Geometry of the Exponential Family

- A statistical model can be modeled as a manifold of distributions by introducing an affine chart in p
- The tangent space in p is defined by $\mathrm{T}_p = \{v : \mathbb{E}_p[v] = 0\}$
- Since $\nabla \mathbb{E}_\theta[f] = \mathrm{Cov}_\theta(f, T)$, the steepest direction is $f - \mathbb{E}_\theta[f]$

Geometry of the Exponential Family

- A statistical model can be modeled as a manifold of distributions by introducing an affine chart in p
- The tangent space in p is defined by $\mathrm{T}_p = \{v : \mathbb{E}_p[v] = 0\}$
- Since $\nabla \mathbb{E}_\theta[f] = \mathrm{Cov}_\theta(f, T)$, the steepest direction is $f - \mathbb{E}_\theta[f]$



- If $f \notin \mathrm{T}_p$, we take the projection $\hat{f}$

Geometry of the Exponential Family

- In case of a finite sample space $\mathcal{X}$

$$\mathsf{T}_\theta = \left\{ v : v = \sum_{i=1}^{k} a_i (T_i(x) - \mathbb{E}_\theta[T_i]), a_i \in \mathbb{R} \right\}$$

and

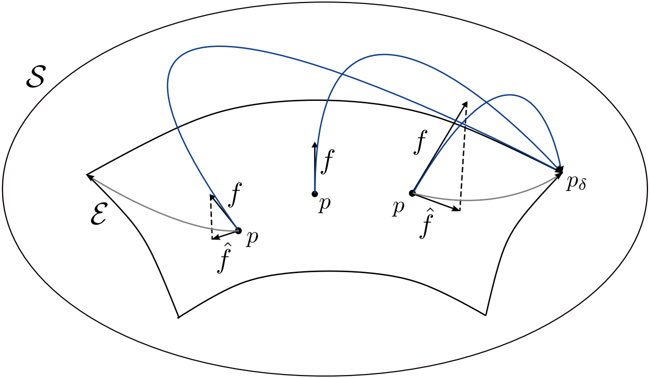
$$\hat{f} = \sum_{i=1}^{k} \hat{a}_i (T_i(x) - \mathbb{E}_\theta[T_i])$$

- Since $f - \hat{f} \perp \mathsf{T}_\theta$ follows that $\mathrm{Cov}_\theta(f - \hat{f}_\theta, T) = 0$ and

$$\hat{a} = \frac{\nabla \mathbb{E}_\theta[f]}{\nabla^2 \psi(\theta)} = \frac{\mathrm{Cov}_\theta(f, T)}{\mathrm{Cov}_\theta(T_i, T_j)}$$

By taking projection of f onto T_p, we obtained the natural gradient, i.e., the gradient evaluated w.r.t. the Fisher information metric

Geometry of the Exponential Family



- If $f \notin \mathrm{T}_p$, the projection $\hat{f}$ may vanish, and local minima appear

Pseudo-Boolean Optimization

- We use the harmonic encoding $\{+1, -1\}$ for binary variables

$$-1^0 = +1 \qquad\qquad -1^1 = -1$$

- A pseudo-Boolean function f is a real-valued map

$$f(x) : \Omega = \{+1, -1\}^n \to \mathbb{R}$$

- Any f can be expanded uniquely as square free polynomial

$$f(x) = \sum_{\alpha \in L} c_\alpha x^\alpha,$$

by employing a multi-index notation, $\alpha = (\alpha_1, \dots, \alpha_n) \in \{0, 1\}^n$

Pseudo-Boolean Optimization

- We use the harmonic encoding $\{+1,-1\}$ for binary variables

$$-1^0 = +1 \qquad -1^1 = -1$$

- A pseudo-Boolean function f is a real-valued map

$$f(x) : \Omega = \{+1,-1\}^n \to \mathbb{R}$$

- Any f can be expanded uniquely as square free polynomial

$$f(x) = \sum_{\alpha \in L} c_\alpha x^\alpha,$$

by employing a multi-index notation, $\alpha = (\alpha_1, \dots, \alpha_n) \in \{0,1\}^n$

- Pseudo-Boolean functions appear in
  - Statistical physics (spin-glass problems)
  - Theoretical computer science (max sat)
  - Machine learning (feature selection, clustering, ranking)
  - Graph theory (max cut)

Expected Fitness Landscape Analysis

Theorem

Consider the stochastic relaxation based on the exponential family $\mathcal{E}$

(i) p_θ in $\mathcal{E}$ is stationary if and only if $\mathrm{Cov}_\theta(f, X^\alpha) = 0$ for all α in M

(ii) if f can be expressed as a linear combination of the sufficient statistics of $\mathcal{E}$, i.e., $f \in \mathrm{Span}\{T_1, \dots, T_k\}$

1. $\nabla \mathbb{E}_\theta[f]$ never vanishes
2. $\mathbb{E}_\eta[f]$ is a linear function in the η parameters

Expected Fitness Landscape Analysis

Theorem

Consider the stochastic relaxation based on the exponential family $\mathcal{E}$

(i) p_θ in $\mathcal{E}$ is stationary if and only if $\mathrm{Cov}_\theta(f, X^\alpha) = 0$ for all α in M

(ii) if f can be expressed as a linear combination of the sufficient statistics of $\mathcal{E}$, i.e., $f \in \mathrm{Span}\{T_1, \ldots, T_k\}$

1. $\nabla \mathbb{E}_\theta[f]$ never vanishes
2. $\mathbb{E}_\eta[f]$ is a linear function in the η parameters

Theorem

If the main effects appear among the sufficient statistics of $\mathcal{E}$, i.e., $\{X_i\}_{i=1}^n \subset \{X^\alpha\}_{\alpha \in M}$, then there exists a sequence of distributions $\{p(x;\theta_t)\}_{t \geq 1}$ in $\mathcal{E}$ such that $\lim_{t\to\infty} p(x;\theta_t) = q$ and $\mathbb{E}_q[f] = \min f$

Expected Fitness Landscape Analysis

Theorem

Consider the stochastic relaxation based on the exponential family $\mathcal{E}$

(i) p_θ in $\mathcal{E}$ is stationary if and only if $\mathrm{Cov}_\theta(f, X^\alpha) = 0$ for all α in M

(ii) if f can be expressed as a linear combination of the sufficient statistics of $\mathcal{E}$, i.e., $f \in \mathrm{Span}\{T_1, \ldots, T_k\}$

1. $\nabla \mathbb{E}_\theta[f]$ never vanishes
2. $\mathbb{E}_\eta[f]$ is a linear function in the η parameters

Theorem

If the main effects appear among the sufficient statistics of $\mathcal{E}$, i.e., $\{X_i\}_{i=1}^n \subset \{X^\alpha\}_{\alpha \in M}$, then there exists a sequence of distributions $\{p(x;\theta_t)\}_{t \geq 1}$ in $\mathcal{E}$ such that $\lim_{t \to \infty} p(x;\theta_t) = q$ and $\mathbb{E}_q[f] = \min f$

Theorem: The Pringles® theorem

Any stationary point of $\mathbb{E}_\theta[f]$ in $\mathcal{E}$ is a saddle point

Stochastic Natural Gradient Descent

- The natural gradient w.r.t. the Fisher information metric is

$$\tilde{\nabla}\mathbb{E}_\theta[f] = \nabla E_\theta[f] I^{-1}(\theta)$$

- The natural gradient is invariant w.r.t. the parametrization and has better convergence properties

Stochastic Natural Gradient Descent

- The natural gradient w.r.t. the Fisher information metric is

$$\tilde{\nabla}\mathbb{E}_\theta[f] = \nabla E_\theta[f] I^{-1}(\theta)$$

- The natural gradient is invariant w.r.t. the parametrization and has better convergence properties
- We can evaluate $\tilde{\nabla}\mathbb{E}_\theta[f]$ by estimating covariances, and explicitly update the model parameters

$$\theta^{t+1} := \theta^t - \gamma \tilde{\nabla}\hat{\mathbb{E}}_\theta[f]$$

Stochastic Natural Gradient Descent

- The natural gradient w.r.t. the Fisher information metric is

$$\tilde{\nabla}\mathbb{E}_\theta[f] = \nabla E_\theta[f] I^{-1}(\theta)$$

- The natural gradient is invariant w.r.t. the parametrization and has better convergence properties
- We can evaluate $\tilde{\nabla}\mathbb{E}_\theta[f]$ by estimating covariances, and explicitly update the model parameters

$$\theta^{t+1} := \theta^t - \gamma \tilde{\nabla}\hat{\mathbb{E}}_\theta[f]$$

Algorithm SNGD(P, γ)

```
Generate a sample 𝒫^0 of size P
t := 0 and θ^0 := 0
repeat
    Evaluate empirical Cov(f, X^α) and Cov(X^α, X^β) from 𝒫^t
    θ^{t+1} := θ^t − γ ∇̃Ê_θ[f]
    Generate 𝒫^{t+1} by sampling P points from p_{θ^{t+1}} with the Gibbs sampler
    t := t + 1
until convergence
```

Stochastic Natural Gradient Descent

- In a single generation approach, we evaluate the gradient once
- Sample $p(x;\theta^1)$ with the Gibbs sampler and a cooling scheme

Stochastic Natural Gradient Descent

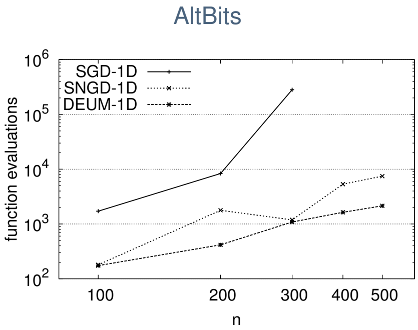
- In a single generation approach, we evaluate the gradient once
- Sample $p(x;\theta^1)$ with the Gibbs sampler and a cooling scheme

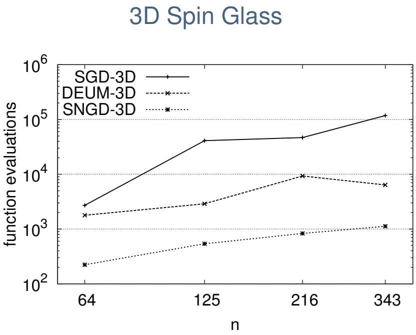
Algorithm GIBBS SAMPLER(p, c, γ)

```
Randomly choose x = (x_1, ..., x_n)
r := 0
repeat
    Set x^tmp := x
    for i ← 1 to n do
        r := r + 1
        T := 1/cr
        Sample x_i from p_i(x_i | x_\i ; θ_i ; T)
    end for
until x^tmp = x or T < γ
return x
```

- Such approach is successful if all interactions of f are captured by the model, i.e., the Gibbs distribution is included in $\mathcal{E}$

SNGD: Experimental Results





L. Malagò, M. Matteucci, and G. Pistone. Stochastic natural gradient descent by estimation of empirical covariances. In *Evolutionary Computation (CEC), 2011 IEEE Congress on*, pages 949 –956, june 2011.

L. Malagò, M. Matteucci, and G. Pistone. Optimization of pseudo-boolean functions by stochastic natural gradient descent. In *MIC 2011, 9th Metaheuristics International Conference*, july 2011.

Sparse Model Selection

- We apply ℓ_1-regularized methods for high-dimensional sparse model selection (Ravikumar et al., 2010)

Sparse Model Selection

- We apply ℓ_1-regularized methods for high-dimensional sparse model selection (Ravikumar et al., 2010)
- Conditional probabilities in the exponential family

$$p_i(x_i|x_{\setminus i};\theta_i) = \frac{1}{1 + \exp\left(-2x_i \sum_{\alpha \in M_i} \theta_{\alpha\setminus i} x^{\alpha\setminus i}\right)}$$

- We reconstruct a sparse neighbourhood for each x_i by solving n different ℓ_1-penalized logistic regression problems

$$\min_{\theta_i} \{\mathcal{L}(\theta_i|\mathcal{P}) + \lambda||\theta_i||_1\}\,, \qquad \lambda = K\sqrt{\frac{\log n}{m}}$$

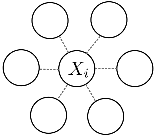
Sparse Model Selection

- We apply ℓ_1-regularized methods for high-dimensional sparse model selection (Ravikumar et al., 2010)
- Conditional probabilities in the exponential family

$$p_i(x_i|x_{\setminus i};\theta_i) = \frac{1}{1+\exp\left(-2x_i \sum_{\alpha\in M_i} \theta_{\alpha\setminus i} x^{\alpha\setminus i}\right)}$$

- We reconstruct a sparse neighbourhood for each x_i by solving n different ℓ_1-penalized logistic regression problems

$$\min_{\theta_i} \{\mathcal{L}(\theta_i|\mathcal{P}) + \lambda||\theta_i||_1\}\,, \qquad \lambda = K\sqrt{\frac{\log n}{m}}$$



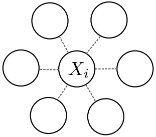
Sparse Model Selection

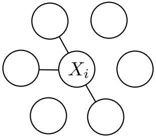
- We apply ℓ_1-regularized methods for high-dimensional sparse model selection (Ravikumar et al., 2010)
- Conditional probabilities in the exponential family

$$p_i(x_i|x_{\setminus i};\theta_i) = \frac{1}{1 + \exp\left(-2x_i \sum_{\alpha \in M_i} \theta_{\alpha\setminus i} x^{\alpha\setminus i}\right)}$$

- We reconstruct a sparse neighbourhood for each x_i by solving n different ℓ_1-penalized logistic regression problems

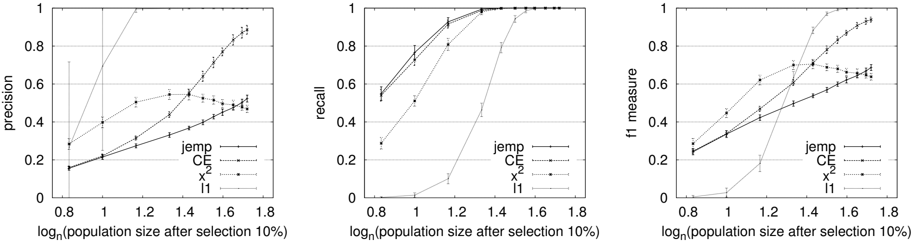
$$\min_{\theta_i} \left\{\mathcal{L}(\theta_i|\mathcal{P}) + \lambda||\theta_i||_1\right\}, \qquad \lambda = K\sqrt{\frac{\log n}{m}}$$





ℓ_1-constrained Model Selection: Experimental Results

2D Spin Glass, n=64



L. Malagò, M. Matteucci, and G. Valentini. Introducing ℓ_1-regularized logistic regression in Markov networks based EDAs. In *Evolutionary Computation (CEC), 2011 IEEE Congress on*, pages 1581 –1588, june 2011.

Markov Fitness Model

- In DEUM (Shakya et al., 2005), p are chosen to be proportional to f

$$p(x) = \frac{f(x)}{\sum_{\Omega} f(x)}$$

Markov Fitness Model

- In DEUM (Shakya et al., 2005), p are chosen to be proportional to f and p belongs to the exponential family

$$p(x) = \frac{f(x)}{\sum_{\Omega} f(x)} \qquad p(x;\theta) = \exp\left\{\sum_{\alpha \in M} \theta_\alpha x^\alpha - \psi(\theta)\right\},$$

Markov Fitness Model

- In DEUM (Shakya et al., 2005), p are chosen to be proportional to f and p belongs to the exponential family

$$p(x) = \frac{f(x)}{\sum_\Omega f(x)} \qquad p(x;\theta) = \exp\left\{\sum_{\alpha \in M} \theta_\alpha x^\alpha - \psi(\theta)\right\},$$

which in particular is satisfied by

$$\ln f(x) = \sum_{\alpha \in M} \theta_\alpha x^\alpha,$$

Markov Fitness Model

- In DEUM (Shakya et al., 2005), p are chosen to be proportional to f and p belongs to the exponential family

$$p(x) = \frac{f(x)}{\sum_\Omega f(x)} \qquad p(x;\theta) = \exp\left\{\sum_{\alpha\in M} \theta_\alpha x^\alpha - \psi(\theta)\right\},$$

which in particular is satisfied by

$$\ln f(x) = \sum_{\alpha\in M} \theta_\alpha x^\alpha,$$

- Parameters are obtained by solving a linear regression problem by least squares

$$\min_{\theta\in\mathbb{R}^k} \left\{\frac{1}{2}\left(\ln f(x) - \sum_{\alpha\in M} \theta_\alpha x^\alpha\right)^2\right\}$$

Linear Regression and Gradient Estimation with Orthogonal Variables

- In DEUM a linear model for $\ln f$ is estimated

$$\ln f(x) = \sum_{\alpha \in M} \theta_\alpha x^\alpha$$

- In the uniform distribution p_0, all X^α are orthogonal, thus regression coefficients can be evaluated as

$$\hat{\theta}_\alpha = \frac{\langle f, x^\alpha \rangle}{\langle x^\alpha, x^\alpha \rangle} = \frac{1}{P} \sum_{\Omega} f x^\alpha = \mathbb{E}_0[f x^\alpha] = c_\alpha$$

Linear Regression and Gradient Estimation with Orthogonal Variables

- In DEUM a linear model for $\ln f$ is estimated

$$\ln f(x) = \sum_{\alpha \in M} \theta_\alpha x^\alpha$$

- In the uniform distribution p_0, all X^α are orthogonal, thus regression coefficients can be evaluated as

$$\hat{\theta}_\alpha = \frac{\langle f, x^\alpha \rangle}{\langle x^\alpha, x^\alpha \rangle} = \frac{1}{P} \sum_\Omega f x^\alpha = \mathbb{E}_0[f x^\alpha] = c_\alpha$$

- In SND gradient components are estimated as

$$\partial_\alpha \mathbb{E}_\theta[f] = \mathrm{Cov}_\theta(f, X^\alpha)$$

- In the uniform distribution p_0, $\mathbb{E}_0[X^\alpha] = 0$, so that and

$$\mathrm{Cov}_0(f, X^\alpha) = \mathrm{E}_0(f X^\alpha) = c_\alpha$$

DEUM, SGD and SNGD

At $t = 0$, in p_0, $\mathbb{E}_0[X^\alpha X^\beta] = 0$, unless $\alpha = \beta$

DEUM

$$\ln f(x) = \sum_{\alpha \in M} \theta_\alpha x^\alpha$$

$$\theta^1 := -\tilde{\nabla} \hat{\mathbb{E}}_0[\ln f]$$

DEUM, SGD and SNGD

At $t = 0$, in p_0, $\mathbb{E}_0[X^\alpha X^\beta] = 0$, unless $\alpha = \beta$

DEUM

$$\ln f(x) = \sum_{\alpha \in M} \theta_\alpha x^\alpha$$

$$\theta^1 := -\tilde{\nabla}\hat{\mathbb{E}}_0[\ln f]$$

SGD

$$\theta^1 := -\gamma \nabla \hat{\mathbb{E}}_0[f]$$

$$f(x) = \sum_{\alpha \in M} \theta_\alpha x^\alpha$$

(under $X^\alpha \perp X^\beta$)

DEUM, SGD and SNGD

At $t = 0$, in p_0, $\mathbb{E}_0[X^\alpha X^\beta] = 0$, unless $\alpha = \beta$

DEUM

$$\ln f(x) = \sum_{\alpha \in M} \theta_\alpha x^\alpha$$

$$\theta^1 := -\tilde{\nabla}\hat{\mathbb{E}}_0[\ln f]$$

SGD

$$\theta^1 := -\gamma \nabla \hat{\mathbb{E}}_0[f]$$

$$f(x) = \sum_{\alpha \in M} \theta_\alpha x^\alpha$$

(under $X^\alpha \perp X^\beta$)

DEUM makes a step in the direction of $\tilde{\nabla}\mathbb{E} \ln f(x)$, and
SNG estimates a linear model for f with orthogonal variables

DEUM, SGD and SNGD

At $t = 0$, in p_0, $\mathbb{E}_0[X^\alpha X^\beta] = 0$, unless $\alpha = \beta$

DEUM

$$\ln f(x) = \sum_{\alpha \in M} \theta_\alpha x^\alpha$$

$$\theta^1 := -\tilde{\nabla}\hat{\mathbb{E}}_0[\ln f]$$

SGD

$$\theta^1 := -\gamma \nabla \hat{\mathbb{E}}_0[f]$$

$$f(x) = \sum_{\alpha \in M} \theta_\alpha x^\alpha$$

(under $X^\alpha \perp X^\beta$)

DEUM makes a step in the direction of $\tilde{\nabla}\mathbb{E} \ln f(x)$, and SNG estimates a linear model for f with orthogonal variables

For any p, SNGD solves $f(x) = \sum_{\alpha \in M} \theta_\alpha x^\alpha$, $\quad \theta^{t+1} := \theta^t - \gamma \tilde{\nabla}\hat{\mathbb{E}}_\theta[f]$

Implications for Model Selection

- Fitness estimation and gradient estimation are strongly connected

Implications for Model Selection

- Fitness estimation and gradient estimation are strongly connected
- Gradient descent can be combined with model selection methods from linear regression
  - Forward stepwise regression
  - LASSO/LAR

Implications for Model Selection

- Fitness estimation and gradient estimation are strongly connected
- Gradient descent can be combined with model selection methods from linear regression
  - Forward stepwise regression
  - LASSO/LAR
- Orthogonality of variables in p_θ allows to test if $\nabla_\alpha \mathbb{E}_\theta[f] \neq 0$ rather then $\tilde{\nabla}_\alpha \mathbb{E}_\theta[f] \neq 0$ (speedup VS accuracy)
- Map $\{X\}_{\alpha \in M}$ to a new set of variables Z, orthogonal in p_θ, and evaluate regular gradients for model selection