Color test

I just want to be

sure

that all my favourite colors

are

being displayed correctly on this

new

device. If not I'll modify them.

Topics at the interface MCMC/ES

ANR Siminole Meeting

Rémi Bardenet

LAL, LRI, University of Paris-Sud XI

February 4th 2011

First attempts at getting the best of both worlds

- In discrete sampling, several authors (see Strens 2003) include random flips and different types of cross-over moves.
- Liang & Wong (2000) 's Evolutionary MC runs parallel chains at different temperatures and proposes cross-over moves.
- Ter Braak (2006) 's differential evolution MC approximates AM by

$$x_{t+1} = x_t + \gamma(x_u - x_v) + \epsilon, \quad u, v \sim \mathcal{U}_{\{1,\ldots,t\}},$$

- Müller & Sbalzarini (2010) present Gaussian Adaptation as a "unifying framework" between (1+1) CMA-ES and the AM algorithm of Haario et al. (2001).

- In the original GaA,

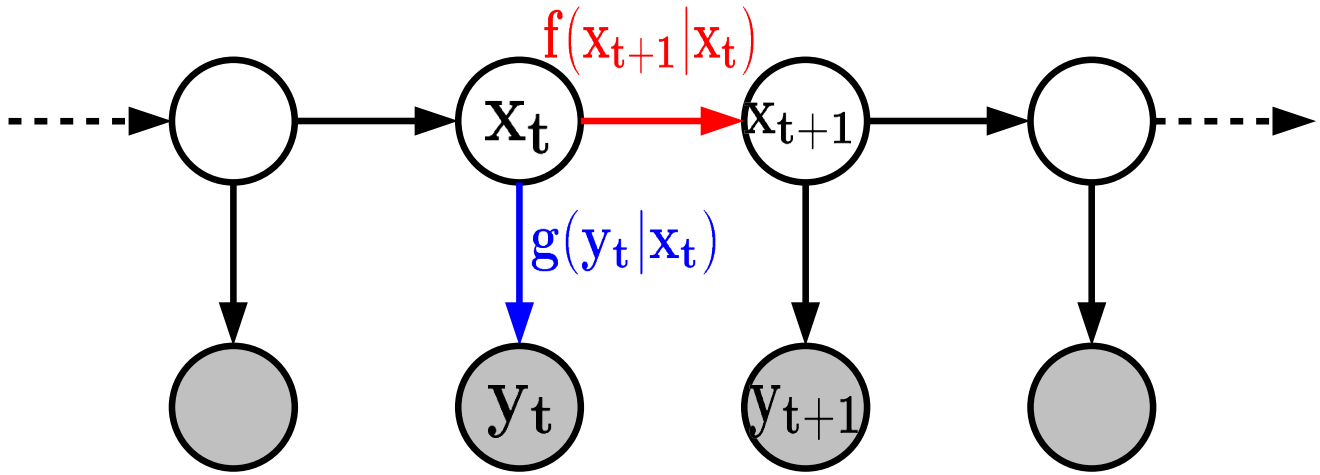
$$\mathbf{C}^{(g+1)} = (1-\gamma_C)\mathbf{C}^{(g)} + \gamma_C(\mathbf{x}^{(g+1)} - \mathbf{x}^{(g)})(\mathbf{x}^{(g+1)} - \mathbf{x}^{(g)})^T.$$

- Differences with CMA-ES include
  - step size adaptation,
  - same covariance update only because $\lambda = \mu = 1$,
  - prior setting of the probability of success.
- When using an MH acceptation rule and a SG-type update for the step size, GaA is very close to Adaptive MCMC.
- Metropolis GaA experimentally behaves comparably to AM.

My 2 cents

- GaA does not satisfy diminishing adaptation.
- Neither does CMA, and it is not desirable, so what is the fundamental difference with sampling ?
- Counterexample by Roberts & Rosenthal is similar to AM, but does not imply that optimization fails.
- What is the role of the learning rate in SGD ?
- Existing SGD with constant learning rate !
- Could mean field-based stochastic approximation be a better common framework ?

A SMC framework for the CEM 1/3 (Zhou et al. 09)



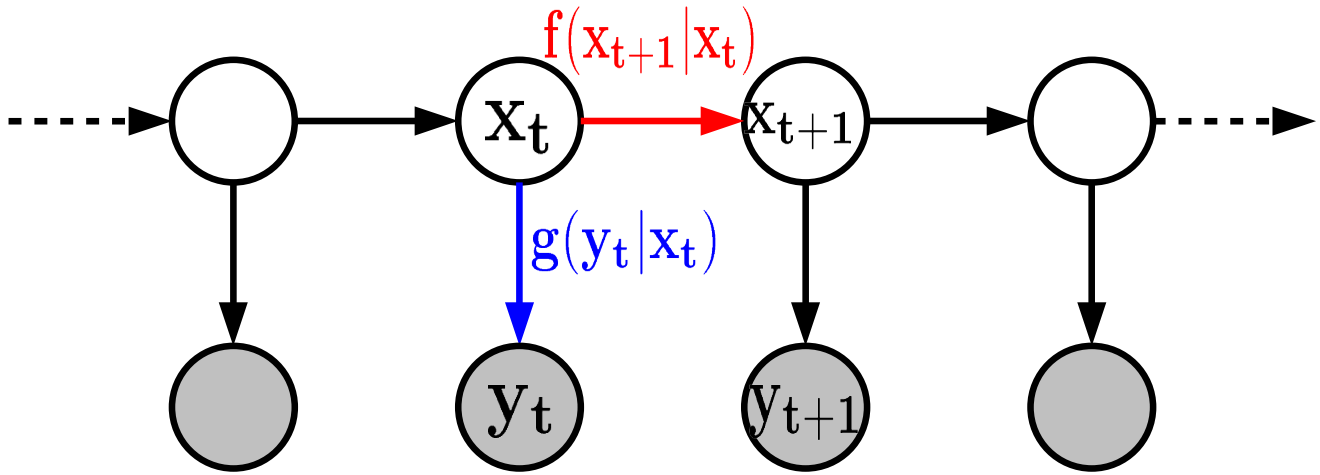
- Goal is to maximize $M_1 \leq H \leq M_2$ over $\mathcal{X}$.
- Assume maximizer x^* is unique.
- Consider the state space model

$$\begin{aligned} x_0 &= x^*, \\ x_{t+1} &= x_t \\ y_t &= H(x_t) - v_t, \quad v_t \sim \varphi. \end{aligned}$$

- Denote $\pi_t(x_t) = p(x_t|y_{0:t})$. Then

$$\hat{\pi}_t(x_t) \propto \varphi(H(x_t) - y_t)\hat{\pi}_{t-1}(x_t)$$

A SMC framework for the CEM 1/3 (Zhou et al. 09)



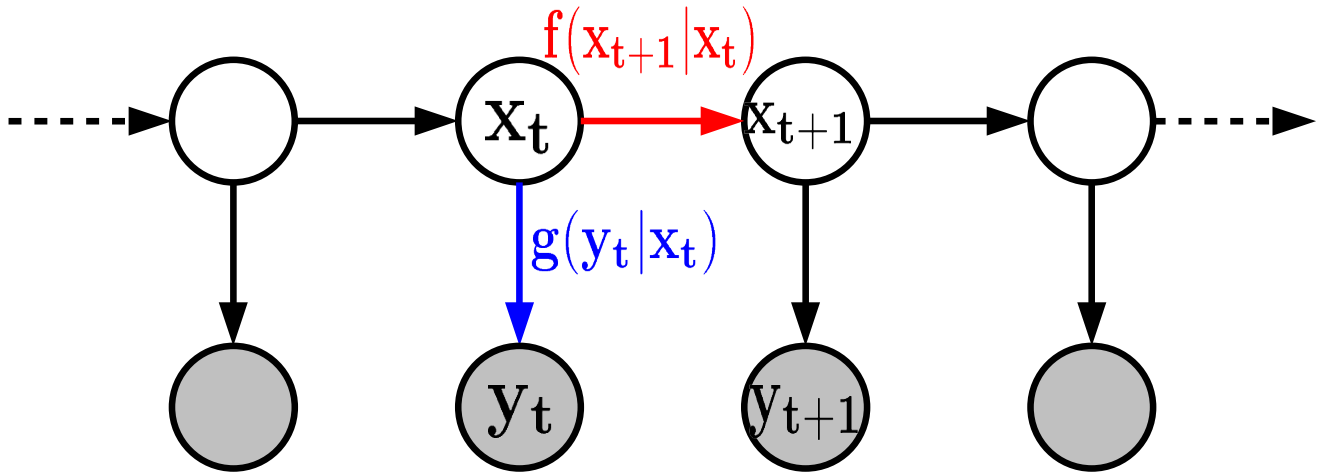
- Goal is to maximize $M_1 \leq H \leq M_2$ over $\mathcal{X}$.
- Assume maximizer $x^\star$ is unique.
- Consider the state space model

$$\begin{aligned} x_0 &= x^*, \\ x_{t+1} &= x_t \\ y_t &= H(x_t) - v_t, \quad v_t \sim \varphi. \end{aligned}$$

- Denote $\pi_t(x_t) = p(x_t|y_{0:t})$. Then

$$\hat{\pi}_t(x_t) \propto \varphi(H(x_t) - y_t)\hat{\pi}_{t-1}(x_t)$$

A SMC framework for the CEM 1/3 (Zhou et al. 09)



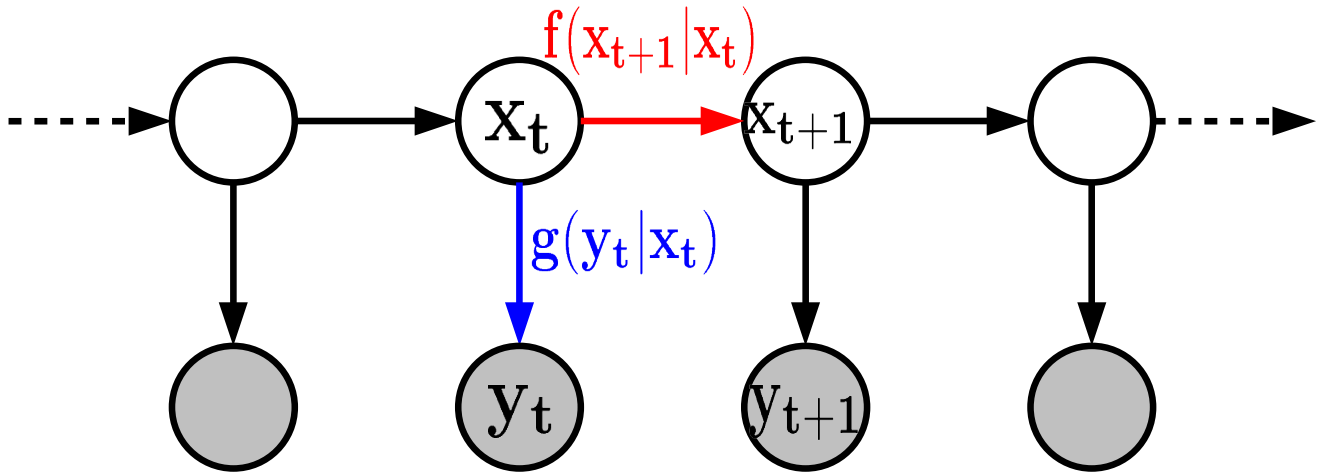
- Goal is to maximize $M_1 \leq H \leq M_2$ over $\mathcal{X}$.
- Assume maximizer $x^\star$ is unique.
- Consider the state space model

$$\begin{aligned} x_0 &= x^*, \\ x_{t+1} &= x_t \\ y_t &= H(x_t) - v_t, \quad v_t \sim \varphi. \end{aligned}$$

- Denote $\pi_t(x_t) = p(x_t|y_{0:t})$. Then

$$\hat{\pi}_t(x_t) \propto \varphi(H(x_t) - y_t)\hat{\pi}_{t-1}(x_t)$$

A SMC framework for the CEM 1/3 (Zhou et al. 09)



- Goal is to maximize $M_1 \leq H \leq M_2$ over $\mathcal{X}$.
- Assume maximizer $x^\star$ is unique.
- Consider the state space model

$$\begin{aligned} x_0 &= x^*, \\ x_{t+1} &= x_t \\ y_t &= H(x_t) - v_t, \quad v_t \sim \varphi. \end{aligned}$$

- Denote $\pi_t(x_t) = p(x_t|y_{0:t})$. Then

$$\hat{\pi}_t(x_t) \propto \varphi(H(x_t) - y_t)\hat{\pi}_{t-1}(x_t)$$

A SMC framework for the CEM 2/3

- Recall

$$\hat{\pi}_t(x_t) \propto \varphi(H(x_t) - y_t)\hat{\pi}_{t-1}(x_t),$$

- If $\hat{\pi}_{t-1}(x) = \sum_i w^{(i)}\delta(x - x_{t-1}^{(i)})$, then

$$\hat{\pi}_t(x_t) \propto \sum_i w^{(i)}\varphi(H(x_{t-1}^{(i)}) - y_t))\delta(x_t = x_{t-1}^{(i)}).$$

- Let y_t be the γ-quantile of the ordered fitnesses of the $(x_{t-1}^{(i)})_i$. Assume $p(y_t|x_t) = \mathcal{U}(0, M_2 - M_1)$, then the filtering update is

$$\hat{\pi}_t(x_t) = \frac{\sum_i 1_{H(x_{t-1}^{(i)}) \geq y_t}\delta(x_t - x_{t-1}^{(i)})}{\sum_i 1_{H(x_{t-1}^{(i)}) \geq y_t}}.$$

A SMC framework for the CEM 2/3

- Recall
$$\hat{\pi}_t(x_t) \propto \varphi(H(x_t) - y_t)\hat{\pi}_{t-1}(x_t),$$

- If $\hat{\pi}_{t-1}(x) = \sum_i w^{(i)}\delta(x - x_{t-1}^{(i)})$, then
$$\hat{\pi}_t(x_t) \propto \sum_i w^{(i)}\varphi(H(x_{t-1}^{(i)}) - y_t))\delta(x_t = x_{t-1}^{(i)}).$$

- Let y_t be the γ-quantile of the ordered fitnesses of the $(x_{t-1}^{(i)})_i$. Assume $p(y_t|x_t) = \mathcal{U}(0, M_2 - M_1)$, then the filtering update is
$$\hat{\pi}_t(x_t) = \frac{\sum_i 1_{H(x_{t-1}^{(i)}) \geq y_t}\delta(x_t - x_{t-1}^{(i)})}{\sum_i 1_{H(x_{t-1}^{(i)}) \geq y_t}}.$$

- Recall
$$\hat{\pi}_t(x_t) \propto \varphi(H(x_t) - y_t)\hat{\pi}_{t-1}(x_t),$$

- If $\hat{\pi}_{t-1}(x) = \sum_i w^{(i)}\delta(x - x_{t-1}^{(i)})$, then
$$\hat{\pi}_t(x_t) \propto \sum_i w^{(i)}\varphi(H(x_{t-1}^{(i)}) - y_t))\delta(x_t = x_{t-1}^{(i)}).$$

- Let y_t be the γ-quantile of the ordered fitnesses of the $(x_{t-1}^{(i)})_i$. Assume $p(y_t|x_t) = \mathcal{U}(0, M_2 - M_1)$, then the filtering update is
$$\hat{\pi}_t(x_t) = \frac{\sum_i 1_{H(x_{t-1}^{(i)}) \geq y_t}\delta(x_t - x_{t-1}^{(i)})}{\sum_i 1_{H(x_{t-1}^{(i)}) \geq y_t}}.$$

A SMC framework for the CEM 3/3

- Recall

$$\hat{\pi}_t(x_t) = \frac{\sum_i 1_{H(x_{t-1}^{(i)}) \geq y_t} \delta(x_t - x_{t-1}^{(i)})}{\sum_i 1_{H(x_{t-1}^{(i)}) \geq y_t}}.$$

- Then project $\hat{\pi}_t(x_t)$ onto the space of Gaussians by minimizing the KL divergence, giving you $\tilde{\pi}_t(x_t)$.
- $\tilde{\pi}_t(x_t)$ is a Gaussian with mean the selected sample mean and covariance the selected sample covariance.
- Finally resample N points from $\tilde{\pi}_t(x_t)$.

Remarks

- A CMA update would require another projection.
- The definition of $p(y_t|x_t)$ is not very natural.

A SMC framework for the CEM 3/3

- Recall

$$\hat{\pi}_t(x_t) = \frac{\sum_i 1_{H(x_{t-1}^{(i)}) \geq y_t} \delta(x_t - x_{t-1}^{(i)})}{\sum_i 1_{H(x_{t-1}^{(i)}) \geq y_t}}.$$

- Then project $\hat{\pi}_t(x_t)$ onto the space of Gaussians by minimizing the KL divergence, giving you $\tilde{\pi}_t(x_t)$.
- $\tilde{\pi}_t(x_t)$ is a Gaussian with mean the selected sample mean and covariance the selected sample covariance.
- Finally resample N points from $\tilde{\pi}_t(x_t)$.

Remarks

- A CMA update would require another projection.
- The definition of $p(y_t|x_t)$ is not very natural.

- Recall

$$\hat{\pi}_t(x_t) = \frac{\sum_i 1_{H(x_{t-1}^{(i)}) \geq y_t} \delta(x_t - x_{t-1}^{(i)})}{\sum_i 1_{H(x_{t-1}^{(i)}) \geq y_t}}.$$

- Then project $\hat{\pi}_t(x_t)$ onto the space of Gaussians by minimizing the KL divergence, giving you $\tilde{\pi}_t(x_t)$.
- $\tilde{\pi}_t(x_t)$ is a Gaussian with mean the selected sample mean and covariance the selected sample covariance.
- Finally resample N points from $\tilde{\pi}_t(x_t)$.

Remarks

- A CMA update would require another projection.
- The definition of $p(y_t|x_t)$ is not very natural.

- Recall

$$\hat{\pi}_t(x_t) = \frac{\sum_i 1_{H(x_{t-1}^{(i)}) \geq y_t} \delta(x_t - x_{t-1}^{(i)})}{\sum_i 1_{H(x_{t-1}^{(i)}) \geq y_t}}.$$

- Then project $\hat{\pi}_t(x_t)$ onto the space of Gaussians by minimizing the KL divergence, giving you $\tilde{\pi}_t(x_t)$.
- $\tilde{\pi}_t(x_t)$ is a Gaussian with mean the selected sample mean and covariance the selected sample covariance.
- Finally resample N points from $\tilde{\pi}_t(x_t)$.

Remarks

- A CMA update would require another projection.
- The definition of $p(y_t|x_t)$ is not very natural.

- Natural gradients also appeared recently in the MCMC literature (Girolami & Calderhead, 2010).
- Take e.g. Metropolis-adjusted Langevin algorithms of the form

$$\theta_{t+1} = \theta_t + \frac{\varepsilon^2}{2}\nabla_\theta \log \pi(\theta_t) + \varepsilon z^t$$

  followed by an MH acceptation step.
- On a flat manifold with metric tensor G, it becomes:

$$\theta^{t+1} = \theta^t + \frac{\varepsilon^2}{2} G^{-1}(\theta^t)\nabla_\theta \log \pi(\theta^t) + \varepsilon\sqrt{G^{-1}(\theta^t)}z^t.$$

- However, no evolution gradient here !
- Atchadé (2006) 's adaptive drift algorithms might be interesting for optimization problems.

Conclusion

Take-home message

- Surprisingly similar parallel discoveries in Simulation and Optimization.
- New connections are surely waiting to be drawn.

Thanks for your attention !