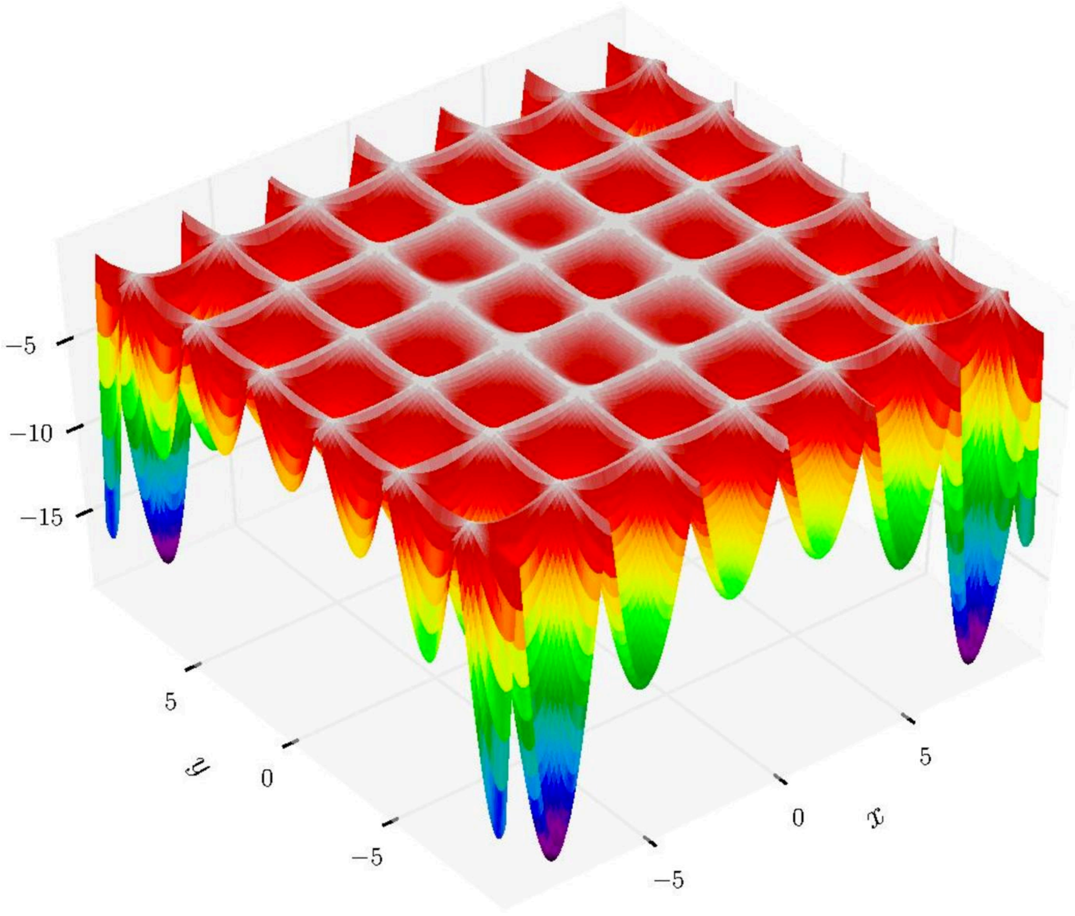
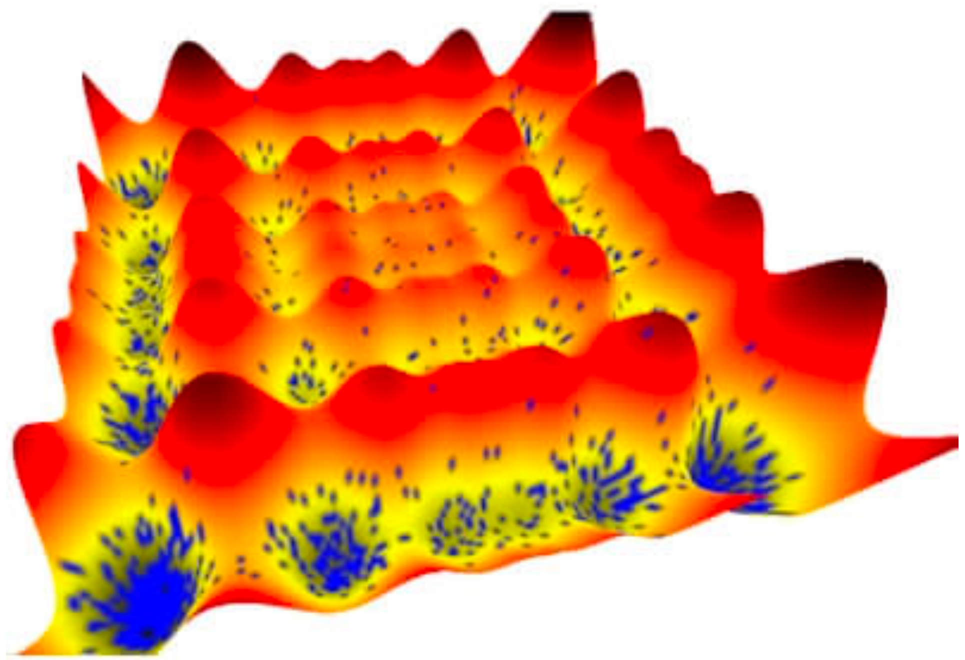
Stochastic Search

UCSD CSE 257
Sicun Gao

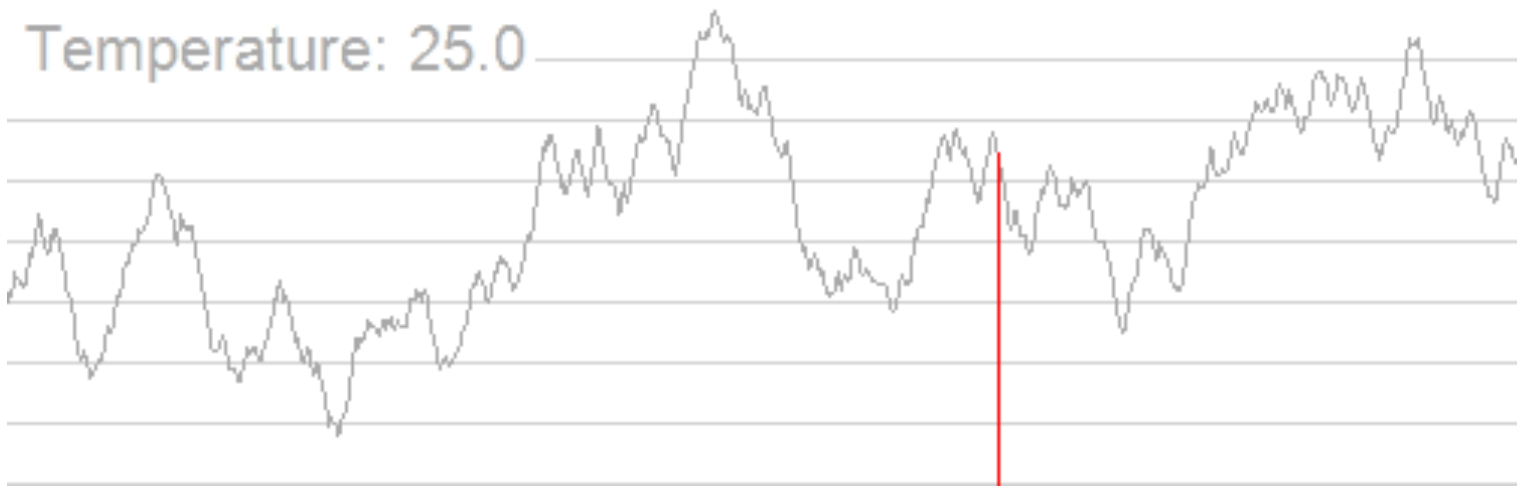
Finding Global Minima





Simulated Annealing

- Random walk, go downhill when you can, but sometimes go uphill to explore
- Gradually settle down (reduce the probability of going uphill over time)



Simulated Annealing

Initialize with random x

Repeat:

Sample a step: $\Delta x \sim P(x)$ some high entropy distribution around x

if $f(x + \Delta x) \geq f(x)$:

With some acceptance probability $x \leftarrow x + \Delta x$

"bad move is sometimes accepted"

else :

$x \leftarrow x + \Delta x$

"good move is always accepted"

Acceptance Probability

- When $f(x + \Delta x) \geq f(x)$, we probabilistically accept the move based on
  - How bad the move is (i.e. $|f(x+\Delta x) - f(x)|$)
  - How much we are interested in exploring
- Accept with probability mass

$$P\left[\text{accept} \,|\, f(x + \Delta x) \geq f(x)\right] = \exp\left(\frac{f(x) - f(x + \Delta x)}{T}\right)$$

$T > 0$: temperature, starting from some large values, decreasing over iterations

Boltzmann/Gibbs distribution $\propto \exp(\frac{-E}{kT})$

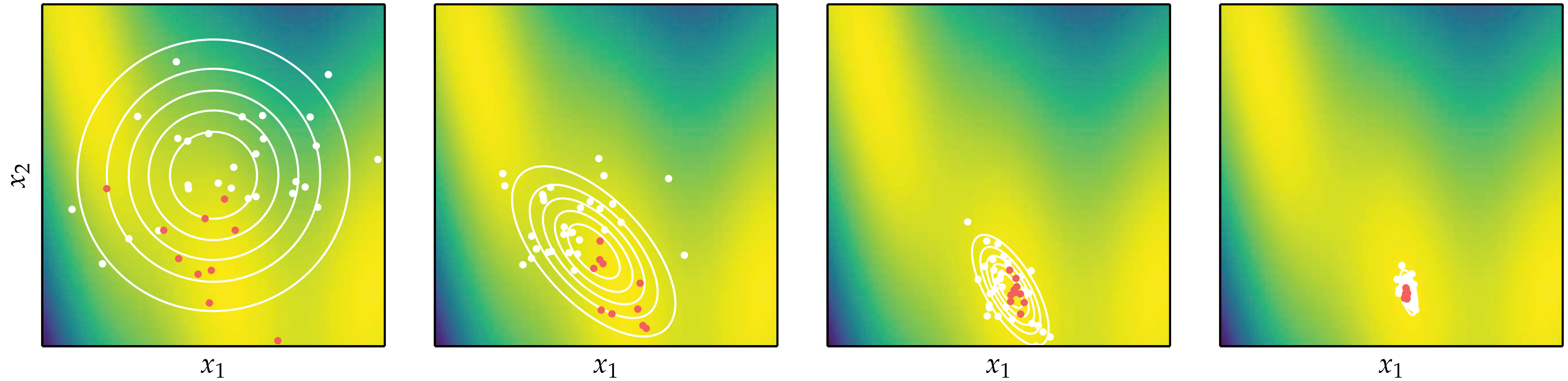
Annealing Schedule (Cooling)

- Fast annealing: $T_k = T/k$
- Exponential annealing: $T_{k+1} = \gamma T_k, \gamma \in (0,1)$
- Log annealing: $T_k = T\log(2)/\log(k+1)$

- Theoretically: asymptotically converge to global minimum
- Practically: curse of dimensionality

Cross Entropy Methods

- Simulated annealing uses a sequence of points
- Cross-entropy methods maintain a sequence of distributions to approach the global minimum



Cross Entropy Methods

- Cross-entropy methods were originally designed for sampling rare events
  - Minimize the divergence between the sampling distribution and the target distribution)
- Adapted to optimization: finding a global optimum is equivalent to sampling a distribution centered around it

Cross Entropy Methods

Start with an initial proposal distribution $p(x)$

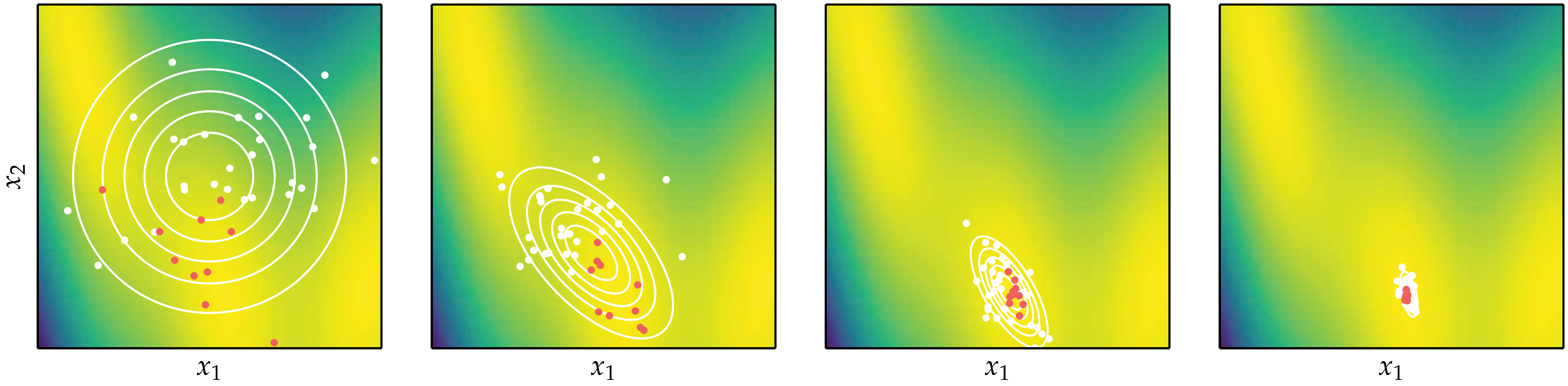
Repeat the following

1. Collect a set A of samples $\sim p(x)$
2. Select a subset of elite samples $E \subseteq A$ (top k samples)
3. Update $p(x)$ to best fit E (do MLE)

Typically $p(x)$ starts from a diagonal Gaussian, updated by:

$$\mu_p \leftarrow \frac{1}{\|E\|} \sum_{x \in E} x \qquad \Sigma_p \leftarrow \frac{1}{\|E\|} \sum_{x \in E} (x - \mu_p)(x - \mu_p)^T$$

Cross Entropy Methods



Pros and Cons?

Search Gradient

- In high-dimensions, it can quickly become very inefficient to randomly sample.

- Ideally we should use gradients again

$$\nabla_\theta \mathbb{E}_{x \sim P_\theta}[f(x)]$$

- So that we can move θ in the directions that improves the expectation.

Search Gradient

$$\nabla_\theta \mathbb{E}_{x \sim P_\theta}[f(x)]$$

$$= \nabla_\theta \int f(x) p_\theta(x) \mathrm{d}x = \int f(x) \Big(\nabla_\theta p_\theta(x) \Big) \mathrm{d}x$$

$$= \int f(x) \Big(\underline{p_\theta(x)} \, \nabla_\theta \log p_\theta(x) \Big) \underline{\mathrm{d}x}$$

$$= \mathbb{E}_{x \sim P_\theta}[f(x) \, \nabla_\theta \log(p_\theta(x))]$$

Search Gradient

- So we just sample $z_1, \ldots, z_k \sim P_\theta$ and evaluate

$$f(z_i)\, \nabla_\theta \log(p_\theta(z_i))$$

and use the average to estimate the expectation

$$\frac{1}{k}\sum_{i=1}^{k} f(z_i)\, \nabla_\theta \log(p_\theta(z_i))$$

$$\longrightarrow \mathbb{E}_{x\sim P_\theta}[f(x)\, \nabla_\theta \log(p_\theta(x))]$$

$$= \nabla_\theta \mathbb{E}_{x\sim P_\theta}[f(x)]$$

Search Gradient

- Overall algorithm

input: f, θ_{init}

repeat

  for $k = 1 \ldots \lambda$ **do**

    draw sample $\mathbf{z}_k \sim \pi(\cdot|\theta)$

    evaluate the fitness $f(\mathbf{z}_k)$

    calculate log-derivatives $\nabla_\theta \log \pi(\mathbf{z}_k|\theta)$

  end

$$\nabla_\theta J \leftarrow \frac{1}{\lambda} \sum_{k=1}^{\lambda} \nabla_\theta \log \pi(\mathbf{z}_k|\theta) \cdot f(\mathbf{z}_k)$$

  $\theta \leftarrow \theta + \eta \cdot \nabla_\theta J$

until *stopping criterion is met*

Limitations of Search Gradient

- Consider n-dimensional normal distribution

$$p_{\mu,\Sigma} = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

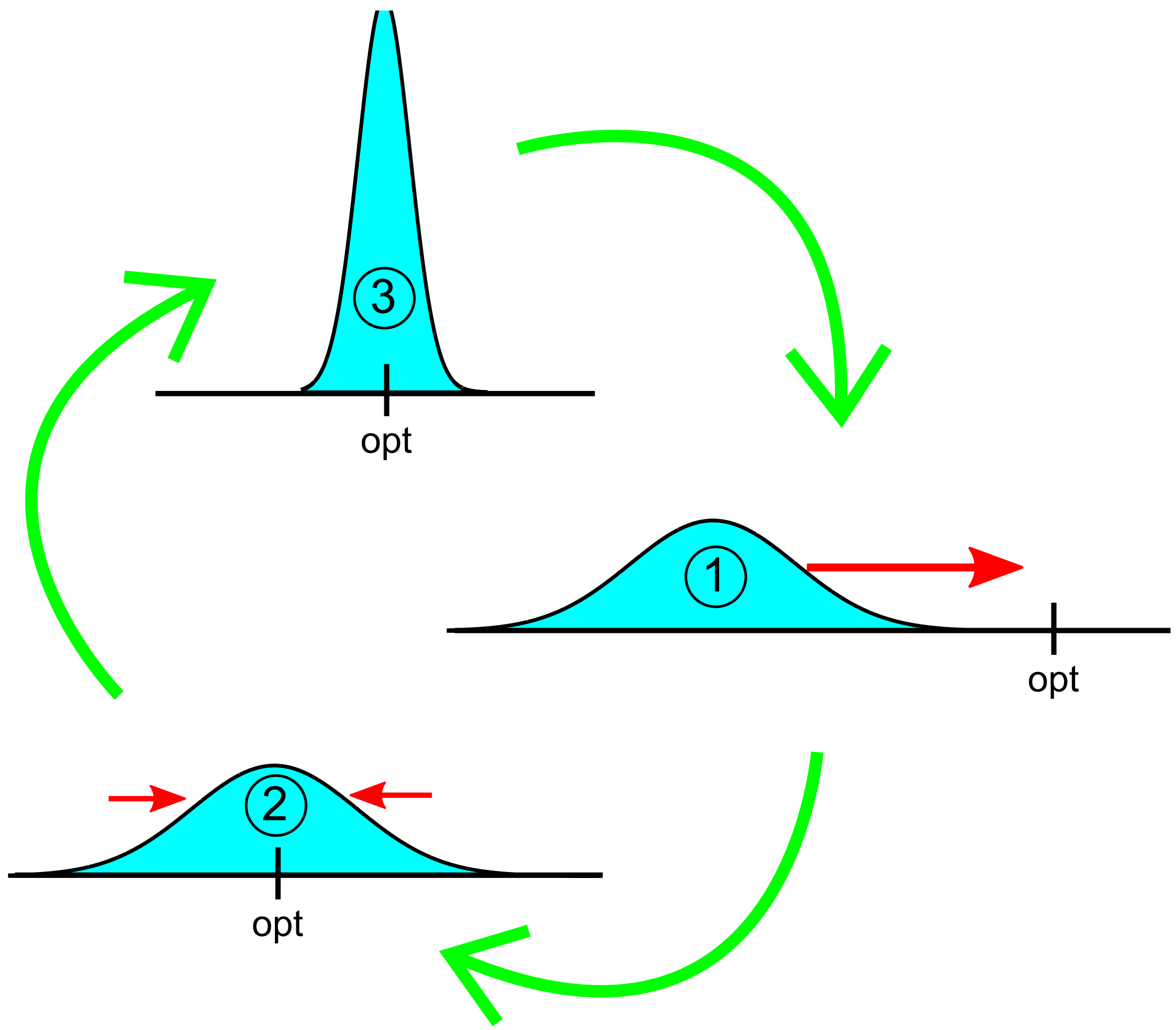
$$\nabla_\mu \log p_{\mu,\Sigma}(z_i) = \Sigma^{-1}(z_i - \mu)$$

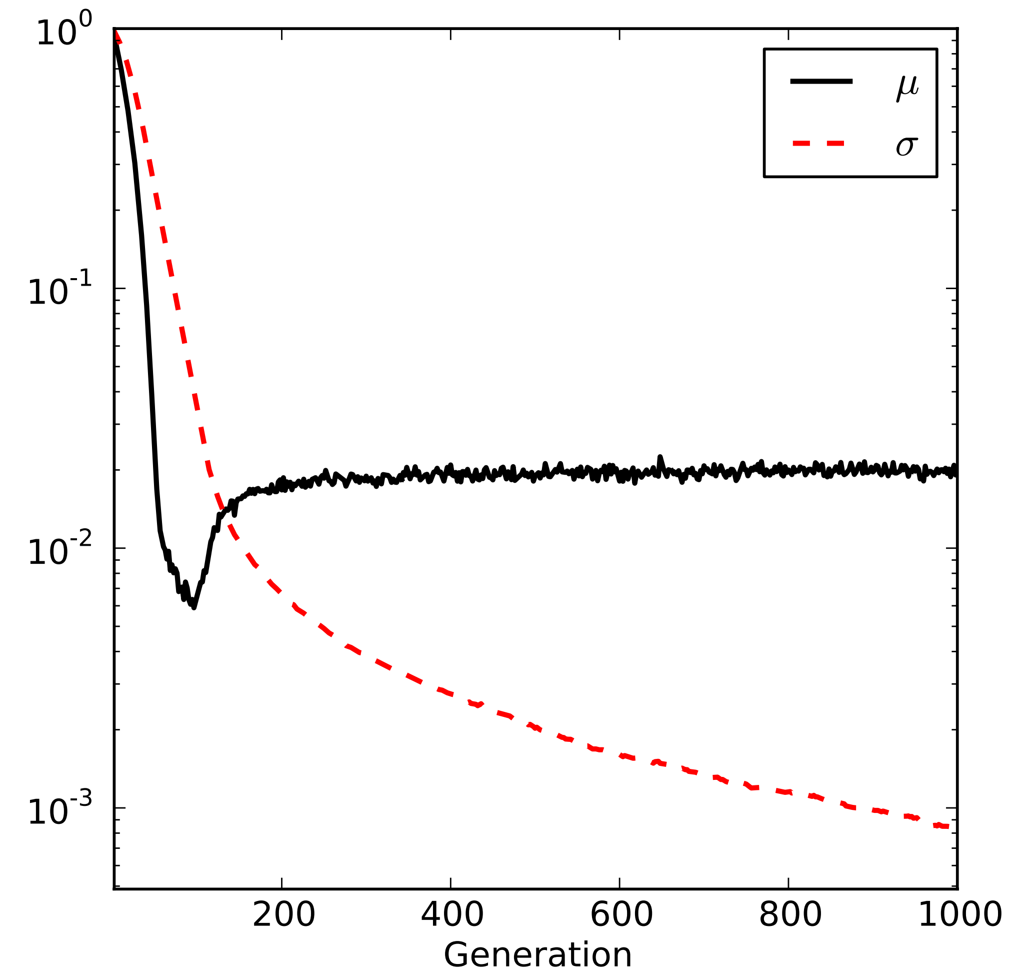
$$\frac{z_i - \mu}{\sigma^2}$$

$$\nabla_\Sigma \log p_{\mu,\Sigma}(z_i) = -\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}(z_i - \mu)(z_i - \mu)^T \Sigma^{-1}$$

$$\frac{(z_i - \mu)^2 - \sigma^2}{\sigma^3}$$

Limitations of Search Gradient





performance on minimizing x^2

"Natural Evolution Strategies" Wierstra et al. JMLR 2014