Week 6, Lectures 2-3

- Trust-region methods
- Global optimization

Cauchy-Point Method (§7.6) EXAM

Approximate the cost function at each iteration x_k as a quadratic problem:

$$f(x_k + p) \simeq f_k + \langle g, p \rangle + \frac{1}{2} \langle p, Bp \rangle$$
$$= m_k(p)$$

The approximation is valid in the ~~t~~ trust region where

$$\|p\|_2 \leq \Delta .$$

At each iteration, we minimize:

$$p_* = \underset{\|p\|_2 \leq \Delta}{\arg\min}\ m_k(p)$$

Theorem 7.1 tells us how to solve for p_* <u>exactly</u>.

$$\left.\begin{array}{l} * \text{ Cauchy Point Method} \\ * \text{ Dogleg Method} \end{array}\right\} \text{Approximate solutions}$$

The idea behind the Cauchy-Point Method:

If Δ is very small, and $\|p\|_2 \leq \Delta$, then the quadratic term in $m_k(p)$ is negligible, and

$$m_k(p) \simeq f_k + \langle g, p \rangle$$

To minimize this, we take

$$p \propto -g.$$

We place p on the trust-region boundary to obtain

$$p_{temp} = -\frac{\Delta}{\|g\|_2} g$$

We next look at $p = \tau\, p_{temp}$ where τ is to be determined, and where τ solves:

$$\tau = \underset{\tau > 0}{\arg\min}\; m_k(\tau\, p_{temp}).$$

We compute τ.

$$m_k(\tau\, p_{temp}) = f_k + \tau \langle g, p_{temp} \rangle + \frac{1}{2}\tau^2 \langle p_{temp}, B\, p_{temp} \rangle$$

We use $p_{temp} = -\left(\Delta / \|g\|_2\right) g$:

$$m_k(\tau p_{temp}) = f_k + \tau \overbrace{\langle g, g\rangle}^{\|g\|_2^2} \left(\frac{-\Delta}{\|g\|_2}\right) + \frac{1}{2}\tau^2 \left(\frac{-\Delta}{\|g\|_2}\right)^2 \langle g, Bg\rangle .$$

Hence :

$$m_k(\tau p_{temp}) = f_k \underbrace{- \tau \Delta \|g\|_2 + \frac{1}{2}\tau^2 \frac{\Delta^2}{\|g\|_2^2} \langle g, Bg\rangle}_{Q(\tau)} .$$

We find the τ which minimizes $Q(\tau)$.

__Two cases__ :

__Case 1__ : $\langle g, Bg\rangle \leq 0$. Then, take $\tau = 1$, to make $Q(\tau)$ as negative as possible. Hence, $p = \tau p_{temp}$, $\tau = 1$, so

$$\boxed{p = -\frac{\Delta}{\|g\|} g}$$

__Case 2__ : $\langle g, Bg\rangle > 0$. Look at $Q'(\tau)$:

$$Q'(\tau) = -\Delta\|g\|_2 + \tau \frac{\Delta^2}{\|g\|_2^2} \langle g, Bg\rangle .$$

To find the min, set $Q'(\tau) = 0$.

Hence:

$$\tau = \frac{\Delta \|g\|_2^3}{\Delta^2 \langle g, Bg \rangle}$$

Or

$$\tau = \frac{\|g\|_2^3}{\Delta \langle g, Bg \rangle}.$$

We require $0 \le \tau \le 1$, to keep p in the trust region. Hence:

$$\tau = \min\left(1, \frac{\|g\|_2^3}{\Delta \langle g, Bg \rangle}\right).$$

Summarizing, the Cauchy point is:

$$\cancel{p_{Cauchy}} \quad \boxed{p_{Cauchy} = -\tau \frac{\Delta}{\|g\|_2} g}$$

where

$$\tau = \begin{cases} 1 & \text{if } \langle g, Bg \rangle \le 0 \\ \min\left(1, \frac{\|g\|_2^3}{\Delta \langle g, Bg \rangle}\right) & \cancel{\text{if}} \text{ o'wise.} \end{cases}$$

∎

Drawback of the Cauchy-Point Method:

$$p^{\text{Cauchy}} \propto -\underline{g}.$$

An obvious choice for $\underline{g}$ is ∇f_k. This gives us back the steepest-descent method (poor convergence, $\|\underline{x}_{k+1} - \underline{x}_*\| \leq C \|\underline{x}_k - \underline{x}_*\|^1$)

$$\boxed{p_* = \underset{\|p\|_2 \leq \Delta}{\arg\min}\ m_k(p)} \qquad (1)$$

A better approximate solution to (1) is given by the Dog-Leg Method (Ch. 8).

If we could momentarily forget that (1) is a constrained minimization, we would compute

$$\nabla_p m_k(p) = 0 \Longrightarrow p = -B^{-1}\underline{g}.$$

Notice, $p = -B^{-1}\underline{g}$ is the Newton descent direction, p^{Newton}. If $\|p^{\text{Newton}}\|_2 \leq \Delta$, then p^{Newton} solves (1).

So the idea of the Dog-Leg Method is to accept p^{Newton} as the solution of (1) in cases when $\|p^{\text{Newton}}\|_2 \leq \Delta$. Otherwise, when $\|p^{\text{Newton}}\|_2 > \Delta$, we construct an approximate solution of (1) using a linear combination of p_{Cauchy} and p_{Newton}:

$$p_* \simeq \overbrace{p_{\text{Cauchy}} + \alpha \left(p_{\text{Newton}} - p_{\text{Cauchy}} \right)}^{p_{\text{approx}}}$$

where $\alpha \in [0,1]$.

We just have to find α such that

$$\|p_{\text{approx}}\|_2 \leq \Delta .$$

In Section 8.2 we show that such an α-value can always be found. Here, we attempt to find an α such that

$$\|p_{\text{approx}}\|_2 = \Delta ,$$

or

$$\|p_{\text{approx}}\|_2^2 = \Delta^2 ,$$

or

$$\langle p_{\text{approx}}, p_{\text{approx}} \rangle = \Delta^2 .$$

Hence, we seek solutions α such that

$$\langle p_{approx}, p_{approx} \rangle = \Delta^2 .$$

$$\Rightarrow \langle p_{Cauchy} + \alpha (p_{Newton} - p_{Cauchy}), \; p_{Cauchy} + \alpha (p_{Newton} - p_{Cauchy}) \rangle = \Delta^2 .$$

$$\Rightarrow \|p_{Cauchy}\|_2^2 + 2\alpha \langle p_{Cauchy}, p_{Newton} - p_{Cauchy} \rangle + \alpha^2 \|p_{Newton} - p_{Cauchy}\|_2^2 = \Delta^2 .$$

Or

$$\alpha^2 \overbrace{\|p_{Newton} - p_{Cauchy}\|_2^2}^{a} + \alpha \, \overbrace{2 \langle p_{Cauchy}, p_{Newton} - p_{Cauchy} \rangle}^{b} + \underbrace{\|p_{Cauchy}\|_2^2 - \Delta^2}_{c} = 0 .$$

Solve for α:

$$\alpha = \frac{-\cancel{2}b \pm \sqrt{\cancel{4}b^2 - \cancel{4}ac}}{\cancel{2}a}$$

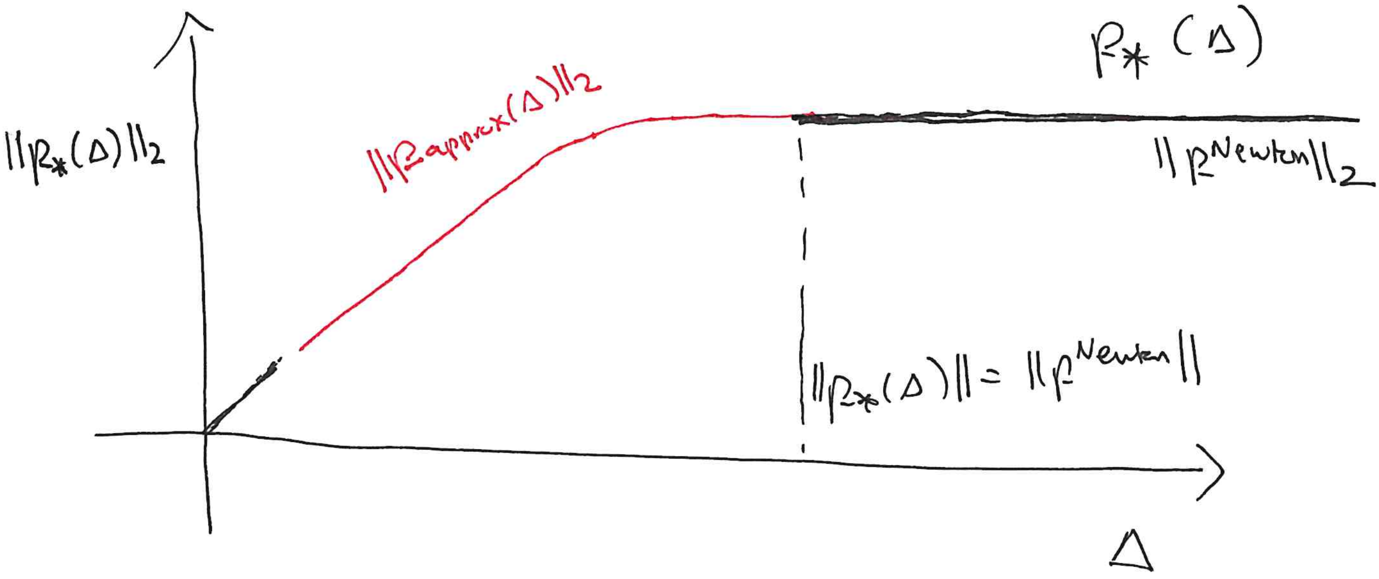
$$\Rightarrow \alpha = \frac{-b \pm \sqrt{b^2 - ac}}{a} .$$

We require $\alpha \in \mathbb{R}$, hence $b^2 - ac \geq 0$.

But notice, $c = \|p_{Cauchy}\|_2^2 - \Delta^2 \leq 0$.

Hence, $b^2 - ac \geq 0$. Hence, a real solution for α exists, giving the Dog-Leg Method.

Reason for the Name:



Key point in favour of the dogleg method:

$$\|\underline{x}_{k+1} - \underline{x}_*\|_2 \leq C \|\underline{x}_k - \underline{x}_*\|^{1+\epsilon}, \quad \epsilon > 0.$$

i.e. superlinear convergence.

Extension to non-positive-definite B (§8.5)

<u>Three Cases</u>

1. When B is positive-definite, we solve (1) in an approximate sense by doing a 2D subspace minimization:

$$p_* = \arg\min m_k(p), \quad \text{subject to}$$
$$p \in \text{Span}(\underline{g}, B^{-1}\underline{g}), \quad \text{and}$$
$$\|p\|_2 \leq \Delta.$$

$$\alpha \underline{g} + \beta B^{-1} \underline{g}$$

2. When B has a zero eigenvalue but no negative eigenvalues, take

$$p_* \simeq p_{\text{Cauchy}}.$$

3. When B has negative eigenvalues, we solve (1) again in an approximate sense by doing another 2D subspace minimization, over

$$p \in \text{Span}(\underline{g}, (B + \alpha I)^{-1}\underline{g}),$$

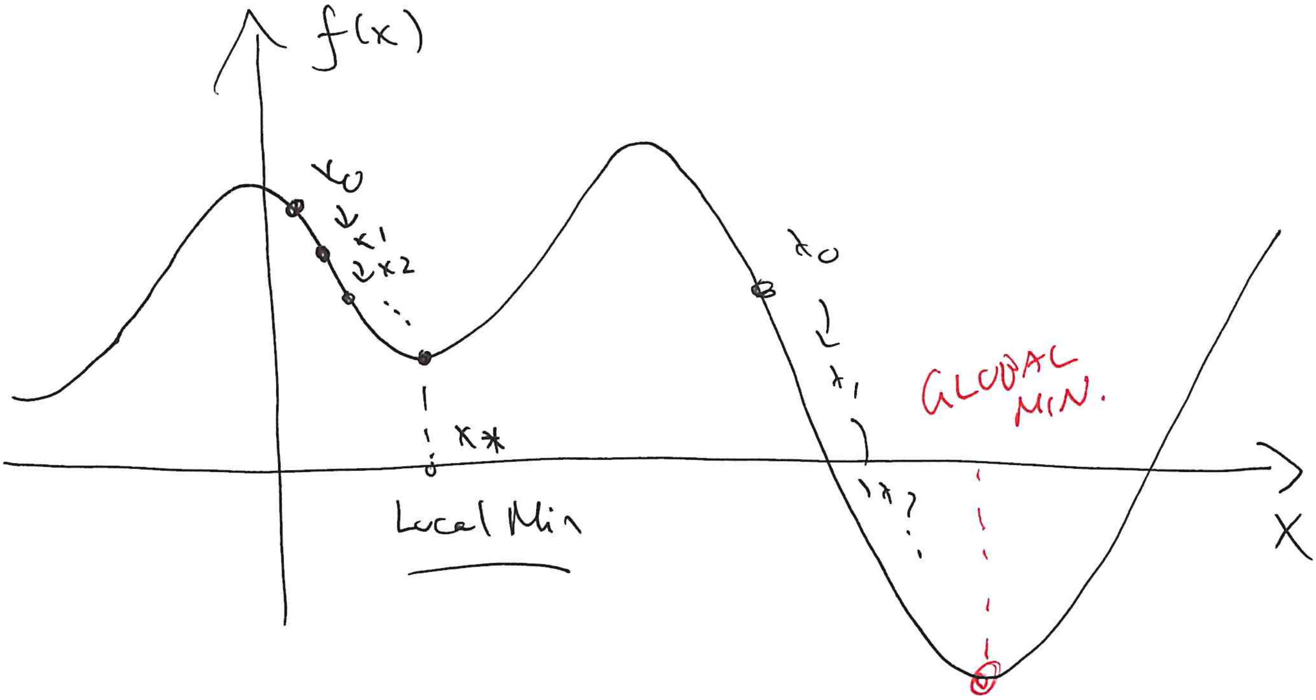
where $\alpha \in (-\lambda_1, -2\lambda_1]$, and where λ_1 is the most negative eigenvalue.

This perturbation makes $B + \alpha \mathbb{I}$ positive-definite.

Week 6, Lecture 3

Global Optimization via Simulated Annealing (Ch. 17).

Motivation — Descent Methods stop when a local min is reached



* We want to find a numerical method that finds the global min, regardless of the value of x_0, the starting-point.

* Global methods like Simulated Annealing, Genetic Algorithms, and Particle Swarm Optimization are robust in another sense as well, as they can deal with non-differentiable cost functions — No differentiation is required.

The idea behind Simulated Annealing (SA) comes from physics. Consider a system (= collection of particles) with n continuous degrees of freedom. A vector $\underline{x} \in \mathbb{R}^n$ therefore describes the state of the system. The energy E of the system is a map from the phase space $\mathbb{R}^n$ to the real numbers:

$$E : \mathbb{R}^n \longrightarrow \mathbb{R} \qquad (1)$$
$$\underline{x} \longmapsto E(\underline{x}) \, .$$

We want to find the probability $d\mathbb{P}$ of finding the system in a small region of phase space of volume $d^n x$, centred at $\underline{x}$:

$$d\mathbb{P} = p(\underline{x})\, d^n x \, . \qquad (2)$$

Here, $p(\underline{x})$ is the probability distribution function, $p(\underline{x}) \geq 0$, and

$$\int_{\mathbb{R}^n} p(\underline{x})\, d^n x = 1. \qquad (3)$$

The energy of the system is thus:

$$\bar{E} = \int_{\mathbb{R}^n} E(x)\, p(x)\, d^n x \; . \qquad (4)$$

The entropy of the system is given by the Boltzmann formula:

$$S = - \int_{\mathbb{R}^n} p(x) \log p(x)\, d^n x \; . \qquad (5)$$

The required p.d.f. $p(x)$ is the one that maximizes the entropy.

$$\tilde{S} = - \int p \log p \, d^n x$$
$$- \beta \left(\int E(x) p(x) d^n x - \bar{E} \right)$$
$$+ \alpha \left(\int p(x) d^n x - 1 \right)$$

Look at small variations in p: $p \to p + \delta p$.
This gives corresponding variations in $\tilde{S}$, $\tilde{S} \to \tilde{S} + \delta \tilde{S}$.

$$\delta \tilde{S} = - \int \underline{\delta (p \log p)}\, d^n x$$
$$- \beta \left(\int E \, \delta p \, d^n x - \cancel{\bar{E}} \right)$$
$$+ \alpha \left(\int \delta p \, d^n x \right) .$$

i.e. $\delta\tilde{S}$ is the change in $\tilde{S}$ after replacing p with $p+\delta p$.

$$\delta(p\log p) = \delta p\log p + p\,\delta\log p$$
$$= \delta p\log p + \cancel{p}\frac{\delta p}{\cancel{p}}$$
$$= \delta p\log p + \delta p.$$

$$\delta\tilde{S} = -\int \delta p\,(\log p+1)\,d^n x$$
$$-\beta\int E\,\delta p\,d^n x \pm \alpha\int\delta p.$$

$$\Rightarrow \delta\tilde{S} = -\int \delta p\underbrace{[\log p+1+\beta E-\alpha]}\,d^n x.$$

$\delta\tilde{S} = 0$ at max entropy.

Hence $[\cdots] = 0$.

$$\Rightarrow \log p + 1 + \beta E - \alpha = 0.$$
$$\Rightarrow \log p = -\beta E + (\alpha-1).$$
$$\Rightarrow p = e^{-\beta E}e^{\alpha-1}.$$

$$1 = \int p(\underline{x})\, d^n x = \left(\int e^{-\beta E(\underline{x})} d^n x \right) e^{\alpha - 1}.$$

This fixes α. Hence:

$$p(\underline{x}) = \frac{e^{-\beta E(\underline{x})}}{\int e^{-\beta E(\underline{x})} d^n x}$$

$$Z = \int e^{-\beta E(\underline{x})} d^n x$$ <u>PARTITION FUNCTION</u>.

$$\boxed{p(\underline{x}) = \frac{e^{-\beta E(\underline{x})}}{Z}}$$ BOLTZMANN DISTRIBUTION

Sub back into S:

$$\tilde{S} = -\int p(\underline{x}) \log p(\underline{x})\, d^n x + \beta(\dots) + \alpha(\dots)$$

$$= -\int p \log p \, d^n x$$

$\log e^{-\beta E}$

$$= -\int \frac{e^{-\beta E(\underline{x})}}{Z} \left[\log \frac{e^{-\beta E}}{Z} \right] d^n x$$

$$= -\int \frac{e^{-\beta E}}{Z} \left[-\beta E - \log Z \right] d^n x.$$

$$\Rightarrow \tilde{S} = \beta \left(\frac{1}{Z} \int e^{-\beta E} E \, d^n x \right) + \log Z \, \underbrace{\frac{1}{Z} \int e^{-\beta E} d^n x}_{=1} .$$

$$= \beta \bar{E} + \log Z .$$

$$\Rightarrow \boxed{S_{max} = \beta \bar{E} + \log Z}$$

$$\frac{\partial \log Z}{\partial \beta} = \frac{1}{Z} \frac{\partial Z}{\partial \beta}$$

$$= \frac{1}{Z} \frac{\partial}{\partial \beta} \int e^{-\beta E(x)} d^n x .$$

$$= -\underbrace{\frac{1}{Z} \int e^{-\beta E} E \, d^n x}$$

$$= - \bar{E} .$$

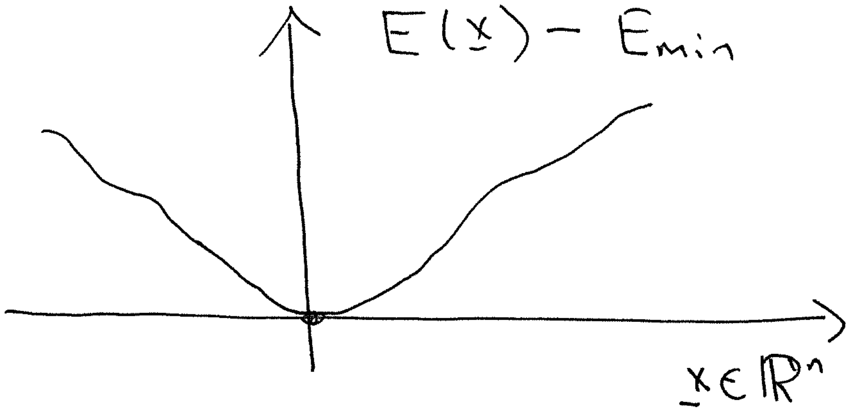
$$\Rightarrow \boxed{\bar{E} = - \frac{\partial \log Z}{\partial \beta}}$$

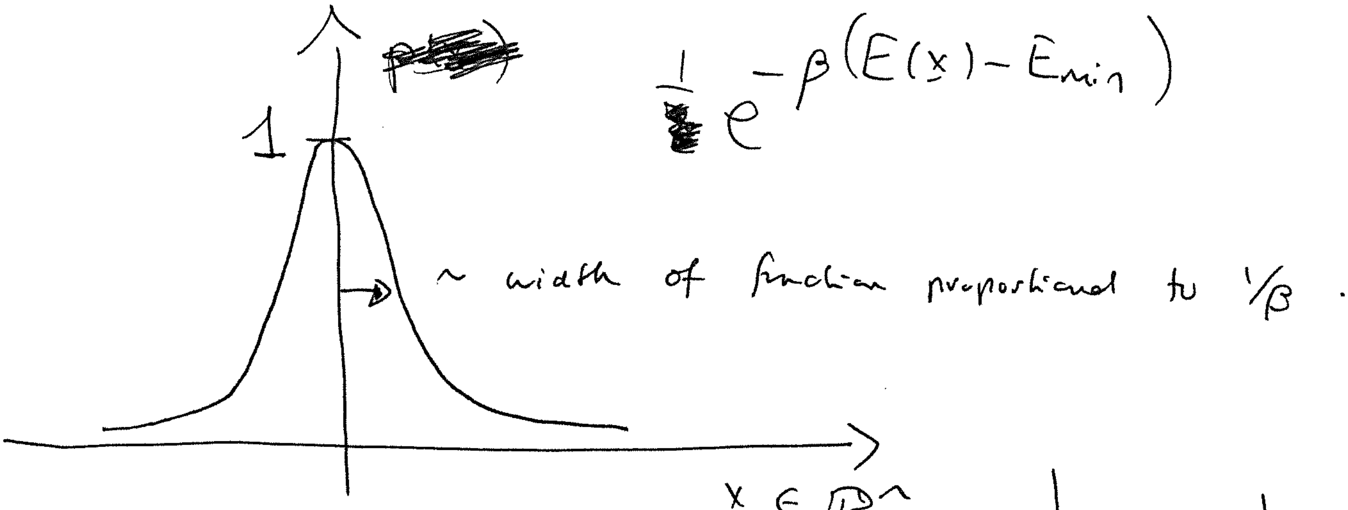
$\beta = \frac{1}{T}$, T = temperature.

$$S_{max} = \beta \bar{E} + \log Z$$

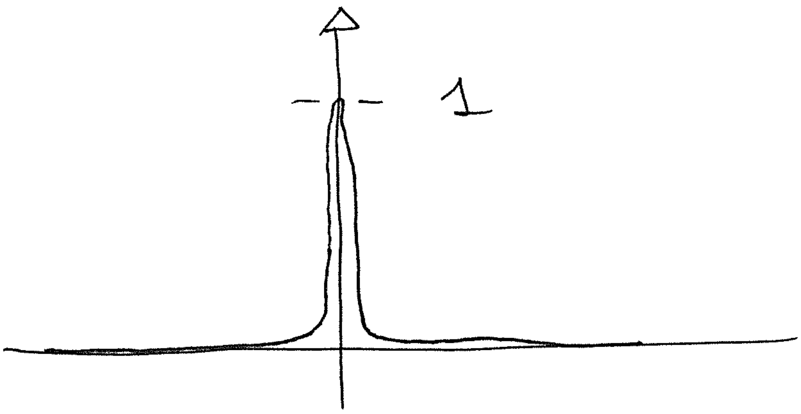
$$\Rightarrow \beta = \frac{\partial S_{max}}{\partial \bar{E}} \Rightarrow \boxed{\frac{1}{T} = \frac{\partial S_{max}}{\partial \bar{E}}}$$

Let's suppose that $E(\underline{x})$ has a global min; call it E_{min}.





$\downarrow$ as $\frac{1}{\beta} \to 0$.



Hence,

$$\frac{e^{-\beta(E(\underline{x}) - E_{min})}}{Z} \longrightarrow \delta(\underline{x} - \underline{x}_*)$$

where $E(\underline{x}_*) = E_{min}$.

This is the idea of the <u>quench</u>: as the system is cooled to $T = 0$, it goes into the minimum-energy state.

So, if we can simulate this quench on a computer, and view $E(\underline{x})$ as the cost function to be minimized, the simulated system reaches a state which minimizes $E(\underline{x})$ — hence, the global min $\underline{x}_*$ of $E(\underline{x})$ can be computed.

This is the <u>simulated-annealing algorithm</u>.

Plan:

- Online lecture on SA algorithm
- Convergence proof.

Next Thursday

- Practical Session — Coding exercises on S.A. (Exercises #6)
- Structure of midterm exam.

HOW TO COMPUTE $\delta\tilde{S}$

$$\tilde{S}[p] = -\int \underline{p \log p}\, d^n x + \cdots$$

$$\tilde{S}[p+\delta p] = -\underline{\int (p+\delta p)\log(p+\delta p)\, d^n x} + \cdots$$

$$\tilde{S}[p+\delta p] - \tilde{S}[p] = \delta\tilde{S}.$$

$$\Rightarrow \delta\tilde{S} = -\int \underline{(p+\delta p)\log(p+\delta p)}\, d^n x + \cdots$$
$$+ \int \underline{p \log p}\, d^n x + \cdots$$

$$= -\int \left[(p+\delta p)\log(p+\delta p) - p\log p\right] d^n x + \cdots$$

$$= -\int \left\{(p+\delta p)\log\left[p\left(1+\frac{\delta p}{p}\right)\right] - p\log p\right\} d^n x + \cdots$$

$$= -\int \left\{(p+\delta p)\left[\log p + \log\left(1+\frac{\delta p}{p}\right)\right] - p\log p\right\} d^n x + \cdots$$

TAYLOR

$$= -\int \left\{(p+\delta p)\left[\log p + \frac{\delta p}{p}\right] - \underline{p\log p}\right\} d^n x + \cdots$$

$$= -\int \left[\delta p \log p + \delta p\right] d^n x + \cdots$$