Motivation.

$$\min_{x \in C} f(x)$$

what if $f(x)$ is non-differentiable, or x is discrete?
what if 'black-box'?

Variational Opt.

$$\min_{x \in C} f(x) \leq \mathop{\mathbb{E}}_{x \sim p(x|\theta)}[f(x)] = J(\theta)$$

Under mild conditions $\min_{x \in C} f(x) = \min_{\theta} J(\theta)$

e.g. if $p(x^*|\theta) = 1$, where $x^* = \arg\min_{x \in C} f(x)$

Is $J(\theta)$ easy to optimize? ~~e.g. differentiable and convex?~~

$$\nabla J(\theta) = \begin{bmatrix} \frac{\partial J(\theta)}{\partial \theta_1} \\ \frac{\partial J(\theta)}{\partial \theta_n} \end{bmatrix} = \frac{\partial J(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta} \int_C f(x)\, p(\theta|x)\, dx$$

$$\frac{\partial}{\partial \theta} \int_C f(x)\, p(\theta|x)\, dx \overset{?}{=} \int_C \frac{\partial}{\partial \theta} f(x)\, p(\theta|x)\, dx$$

related to Leibnitz's rule. weak conds.

$$\nabla J(\theta) = \frac{\partial}{\partial \theta} \int f(x)\, p(x|\theta)\, dx$$

$$= \int \frac{\partial}{\partial \theta} f(x)\, p(x|\theta)\, dx$$

$$= \int f(x) \left(\frac{\partial}{\partial \theta} p(x|\theta) \right) dx$$

$$= \int f(x) \frac{\partial p(x|\theta)}{\partial \theta} \frac{p(x|\theta)}{p(x|\theta)}\, dx$$

$$= \int f(x) \left[\frac{\partial \lg p(x|\theta)}{\partial \theta} \right] \cdot p(x|\theta)\, dx$$

$$= \mathbb{E}_{x \sim p(x|\theta)} \left[f(x) \frac{\partial \lg p(x|\theta)}{\partial \theta} \right]$$

$$= \frac{1}{M} \sum_{i=1}^{M} \left[f(x_i) \nabla_\theta \lg p(x_i|\theta) \right], \quad x_i \sim p(x|\theta)$$

GD.

$$\theta = \theta + \eta \cdot \nabla J(\theta)$$

How to sample $x \sim p(x|\theta)$?

~~what for~~

How to choose $p(x|\theta)$?

So $J(\theta)$ can be differentiable w.r.t. θ.
i.e. $\nabla J(\theta)$ available.

Will $J(\theta)$ be convex in θ ?

THM 1. If $f(x)$ convex, and $p(x|\theta)$ expectation affine then $J(\theta)$ convex in θ.

pf 1. Expectation affine.

$$\mathbb{E}_{x\sim p(x|\theta)}[f(x)] = \mathbb{E}_{z\sim q(z)}[f(\alpha(\theta)z + \beta(\theta))]$$

$$\int f(x)\, p(x|\theta)\, dx = \int q(z)\, f(\alpha(\theta)z + \beta(\theta))\, dz$$

$$=$$

'default' multi-variate normal / Gaussian distribution.

$\mu \in \mathbb{R}^d$. $\Sigma \in \mathbb{R}^{d \times d}$. (covariance matrix a.k.a. mutation matrix for ES.).

To sample efficiently. $x \sim N(x \mid \mu, \Sigma)$

we sample. $s \sim N(s \mid 0, I)$ first $I = \mathrm{diag}(1, \cdots, 1) \in \mathbb{R}^{d \times d}$.

$x = \mu + A^T s$, where $A^T A = \Sigma$.

$$p(x|\theta) = \frac{1}{(\sqrt{2\pi})^d \left|\det(A)\right|} \cdot \frac{\exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right]}{\exp\left(-\frac{1}{2}\| A^{-1}\cdot(x-\mu)\|^2\right)}$$

density of $N(x/\mu, \Sigma)$

$$\log p(x|\theta) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log\det\Sigma - \frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu).$$

$$\nabla_\mu \log p(x|\theta) = \Sigma^{-1}(x-\mu)$$

$$\nabla_\Sigma \log p(x|\theta) = \frac{1}{2}\Sigma^{-1}(x-\mu)(x-\mu)^T\Sigma^{-1} - \frac{1}{2}\Sigma^{-1}$$

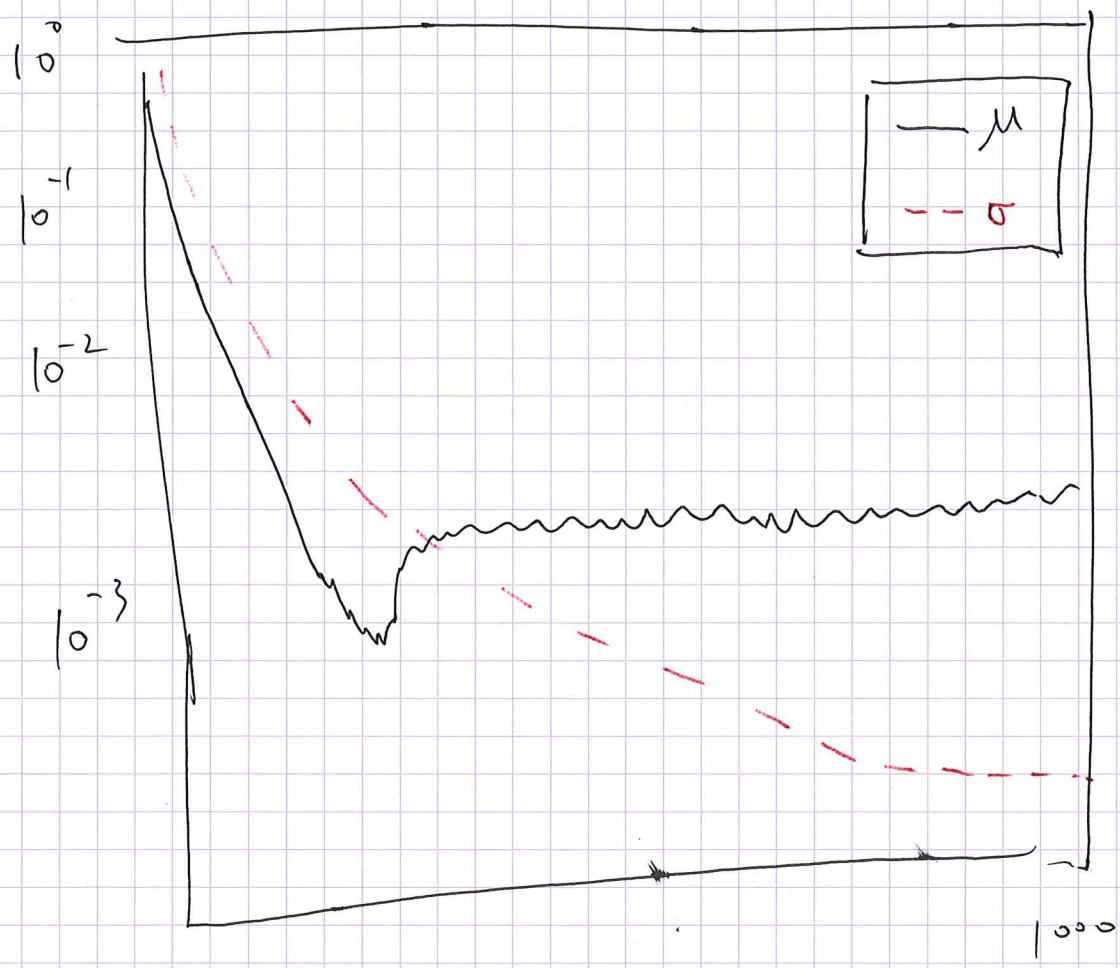
$$\mu \leftarrow \mu - \eta \cdot \nabla_\mu J(\underbrace{\mu, \Sigma}_{\theta})$$

$$\Sigma \leftarrow \Sigma - \eta \cdot \nabla_\Sigma J(\underbrace{\mu, \Sigma}_{\theta})$$

Unstable.

$$\nabla_\mu J = \frac{x-\mu}{\sigma^2}$$

$$\nabla_\sigma J = \frac{(x-\mu)^2 - \sigma^2}{\sigma^3}$$


10^{0}
10^{-1}
10^{-2}
10^{-3}
μ
σ
1000

Natural Gradient

NES = ES + Natural Gradient.

$$\max_{\delta\theta} J(\theta+\delta\theta) \approx J(\theta) + \delta\theta^T \nabla_\theta J$$

$$\text{s.t.} \quad D_{KL}(\theta+\delta\theta \,||\, \theta) = \varepsilon$$

$$D_{KL}(P||Q) = \sum_x P(x) \lg\left(\frac{P(x)}{Q(x)}\right)$$

$$\lim_{\delta\theta\to 0} D_{KL}(\theta+\delta\theta||\theta) = \frac{1}{2}\delta\theta^T \underbrace{E\left[\nabla_\theta \lg P(x|\theta) \nabla_\theta \lg P(x|\theta)^T\right]}_{\text{Fisher infor matrix}} \delta\theta$$

$$\nexists \quad \delta\theta^* = \beta \nabla_\theta J(\theta)$$

$$\delta\theta^* = \beta F^{-1} \nabla_\theta J(\theta)$$

So the natural gradient is

$$\tilde{\nabla}_\theta J(\theta) = F^{-1} \nabla_\theta J(\theta)$$

Fitness shaping

Adaptation sampling (via importance sampling)