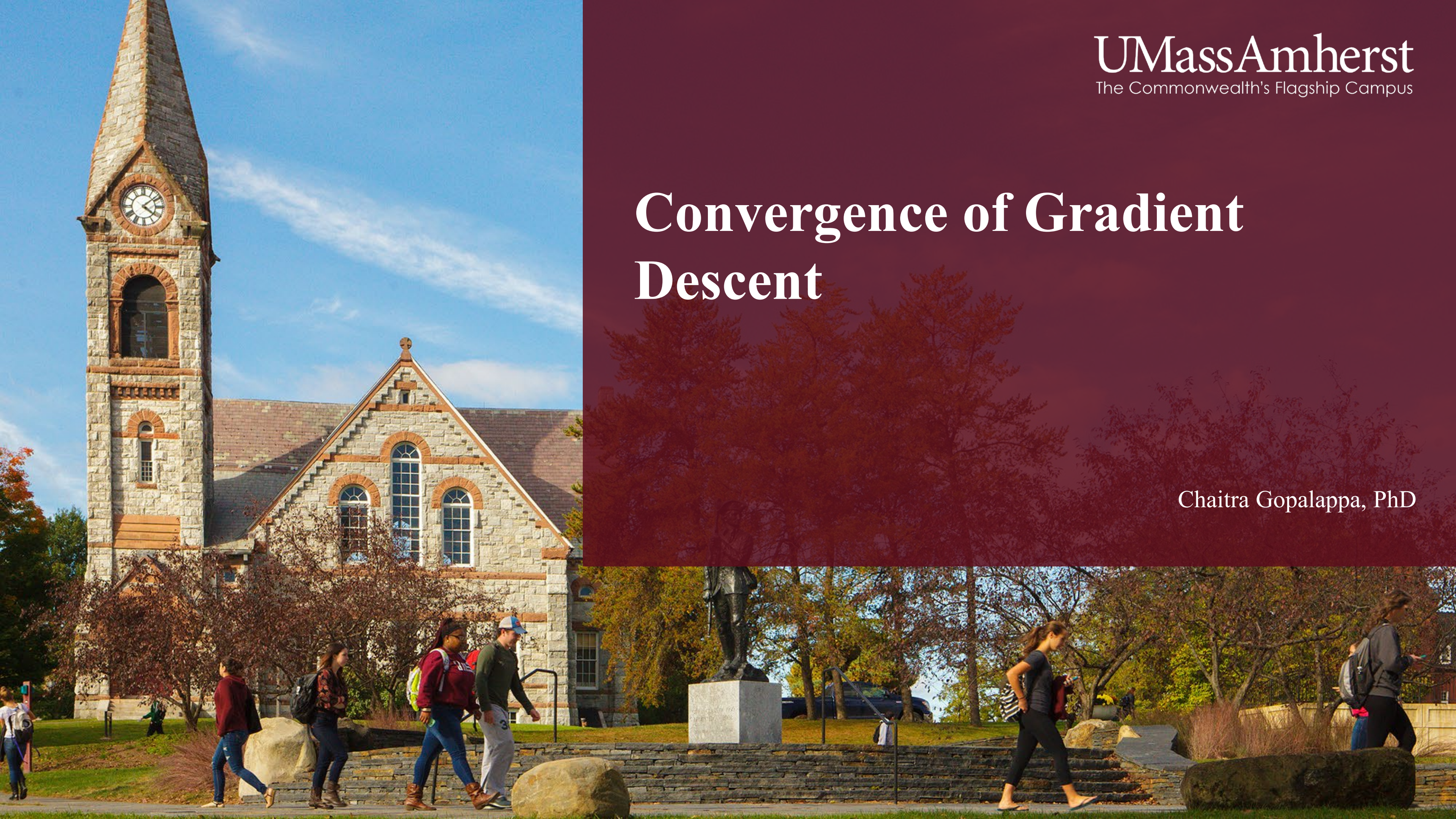


UMassAmherst
The Commonwealth's Flagship Campus

Convergence of Gradient Descent

Chaitra Gopalappa, PhD

Recollect optimality conditions for analytic models: First and second order necessary and sufficient conditions

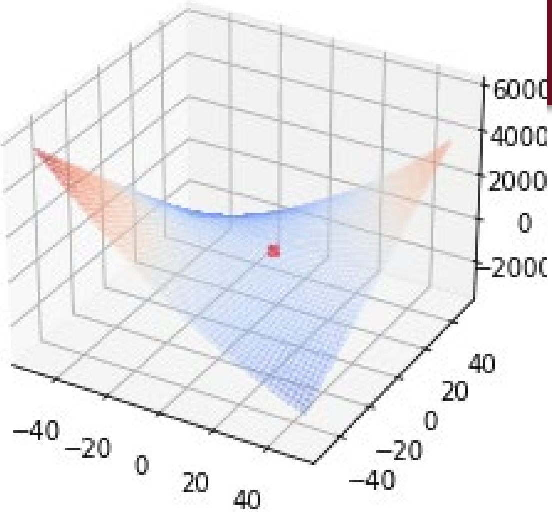
- Suppose $f: R^n \rightarrow R$ is twice differentiable at $\vec{x}*$. If $\vec{x}*$ is a local minimum then
- $\nabla f(\vec{x}*) = 0$ (first-order necessary)
- $H(\vec{x}*)$ is positive semi-definite (second-order necessary)
- $H(\vec{x}*)$ is positive definite (second-order sufficient)

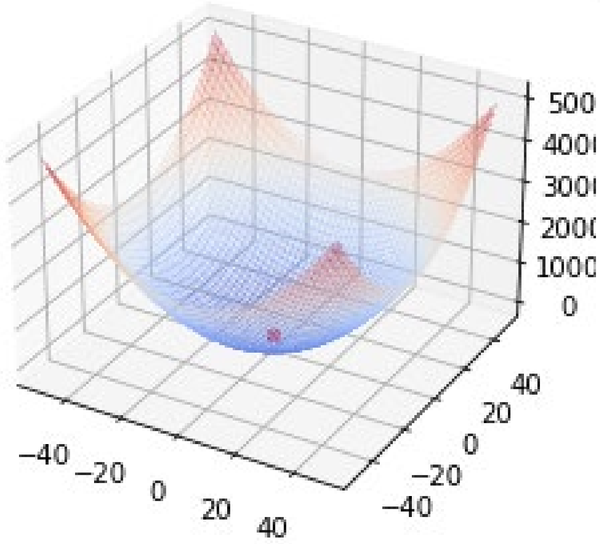
If f is convex and $\vec{x}*$ is a local optima then, $\vec{x}*$ is also a global optima

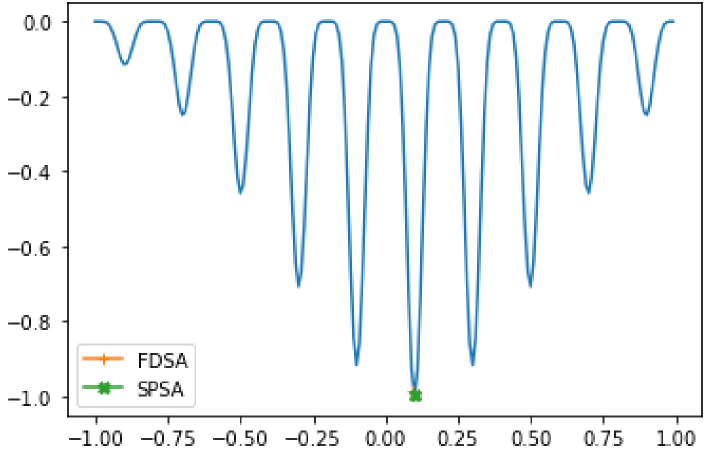
Convergence in search algorithms?

Convergence?

- Does is reach an optimal solution?
- Is it global or local optima?







Recollect: Gradient descent transformation

$$Min\ f(\vec{x})$$
$$\vec{x} \in R^n$$

$$\vec{x}_{m+1} \leftarrow \vec{x}_m - \mu_m \nabla f(\vec{x}_m)$$

Convergence condition- Steepest descent

DEFINITION 9.1 **A Convergent Sequence:** *A sequence* $\{a^p\}_{p=1}^{\infty}$ *is said to converge to a real number* A *iff for any* $\epsilon > 0$, *there exists a positive integer* N *such that for all* $p \geq N$, *we have that*

$$|a^p - A| < \epsilon.$$

Steepest descent transformation:
$\vec{x}_{m+1} \leftarrow \vec{x}_m - \mu_m \nabla f(\vec{x}_m)$;

Does $\vec{x}_m$ converge to $\vec{x}^*{}_m$ (optimal)? Is $\mu_m \nabla f(\vec{x}_m)$ a convergent sequence?
Is following true?

$$|\nabla f(\vec{x}_m) - A| < \epsilon; \;\; A = 0$$

Convergence condition-Steepest descent

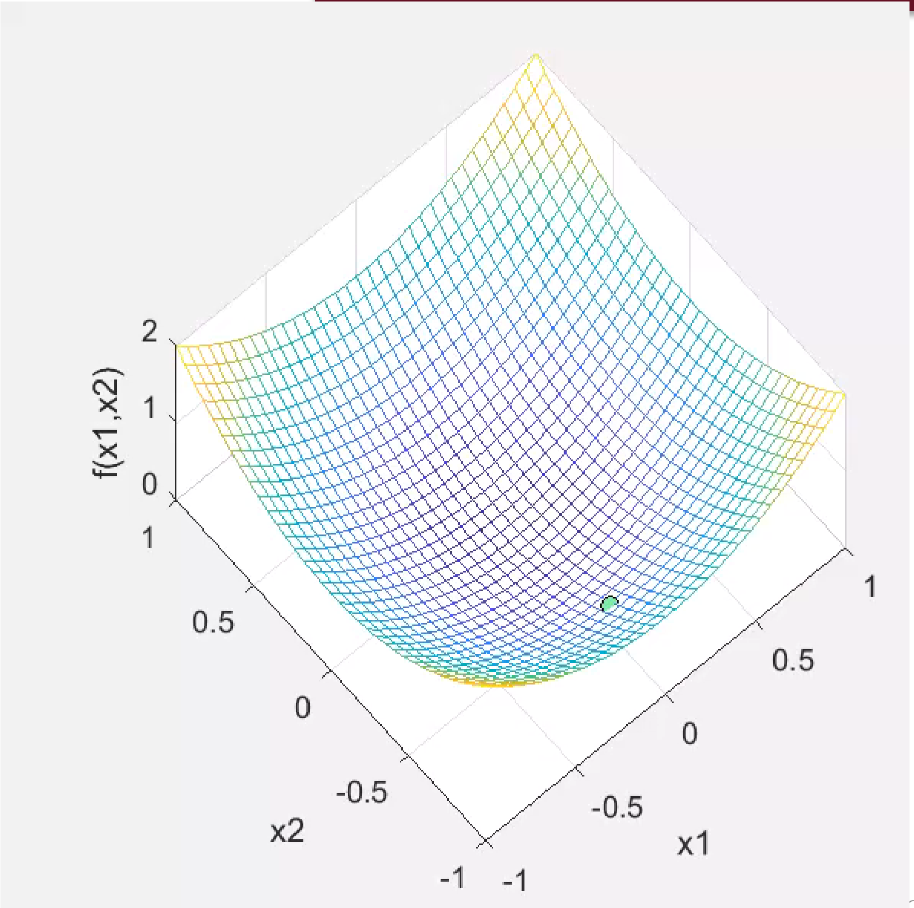
Steepest descent transformation:

$\vec{x}_{m+1} \leftarrow \vec{x}_m - \mu_m \nabla f(\vec{x}_m);$

Does $\vec{x}_m$ converge to $\vec{x}^*{}_m$ (optimal)? Is $\mu_m \nabla f(\vec{x}_m)$ a convergent sequence?

Is following true?

$$|\nabla f(\vec{x}_m) - A| < \epsilon; \quad A = 0$$



Convergence condition- Steepest descent

Steepest descent transformation:
$\vec{x}_{m+1} \leftarrow \vec{x}_m - \mu_m \nabla f(\vec{x}_m)$; Does $\vec{x}_m$ converge to $\vec{x}^*{}_m$ (optimal)?

Theorem: Let $\vec{x}^m$ denote the value in the m^{th} iteration of the steepest-descent approach. If the function f is continuously differentiable, its gradient is Lipschitz continuous, i.e.,

$$||\nabla f(\vec{a_1}) - \nabla f(\vec{a_2})|| \leq L||\vec{a_1} - \vec{a_2}||, \qquad \forall\ \vec{a_1}, \vec{a_2}, \in \Re^k,$$

for some finite $L > 0$, and is bounded below, then for step-size μ < 2/L,

$$\lim_{m \to \infty} \nabla f(\vec{x}^m) = \vec{0}$$

Lipschitz continuous? (for exact gradient)

- $f(x) = 2x \Rightarrow \frac{df}{dx} = 2$ (Yes, LHS of Lipschitz always bounded for any two points of x)
- $2x^2 \Rightarrow \frac{df}{dx} = 2x$ (Depends, LHS of Lipschitz is bounded only if x is not infinity. To ensure Lipschitz continuity bound the function)
- $f(x) = \frac{1}{x}; x \in \{-1,1\} \Rightarrow \frac{df}{dx} = -\frac{1}{x^2}$ (No, LHS of Lipschitz not bounded when $x = 0$)

Stochastic gradient algorithm

- *Transformation*

$$\vec{x}_{m+1} \leftarrow \vec{x}_m - \mu_m Y_m(\vec{x}_m)$$

No change in algorithm; just that it estimates $Y_m(\vec{x}_m)$ instead of $\nabla f(\vec{x}_m)$

$(Y_m(\vec{x}_m) = \nabla f(\vec{x}_m) + noise)$

Convergence with probability 1 (for stochastic approximations)

DEFINITION 9.9 *A sequence* $\{x^k\}_{k=0}^{\infty}$ *of random variables is said to converge to a random number* x_* *with probability 1 if for a given* $\epsilon > 0$ *and a given* $\delta > 0$, *there exists an integer* N *such that*

$$\mathsf{P}\left[|x^k - x_*| < \epsilon\right] > 1 - \delta \text{ for all } k \geq N,$$

$$\text{i.e., } \mathsf{P}\left[\lim_{k\to\infty} x^k = x_*\right] = 1.$$

Example

- Let $X_k = \max\{X_{k-1}, roll\ of\ die\}$
- $P\left[\left|x^k - x^*\right| < \epsilon\right] > 1 - \delta \qquad (1)$
- x^*=6
- K = *random variable defining number of trials to get one success~geometric(p)*
  - $\Pr(K <= k) = 1 - (1-p)^k$ *(cdf)*
- *For* ϵ*=0,* $\delta = 0.1$
  - IF k =10: $\Pr(K <= k) = 1 - (1-p)^k = 1 - \left(1 - \frac{1}{6}\right)^{10} = 0.84 \Rightarrow$ *LHS of (1) <RHS of 1=> k=10 is not a sufficient sample*
  - IF k =20: $\Pr(K <= k) = 1 - (1-p)^k = 1 - \left(1 - \frac{1}{6}\right)^{20} = 0.97 \Rightarrow$ *LHS of (1) >RHS of 1=> k=20 is a sufficient sample*
- *To find what is minimum k*
- Set $1 - (1-p)^k = 1 - \delta \Rightarrow k = 12.6 \sim 13\ samples\ at\ least$

Convergence conditions for stochastic approximations of GD (gradients with noise (ω))

- With probability 1, the sequence $\{f(\vec{x}_{\mathrm{m}})\}_{m=1}^{\infty}$ converges and $\lim_{m\to\infty} \nabla f(\vec{x}_{\mathrm{m}}) \to 0$, if
  - $f(\vec{x}) \geq 0$ *everywhere.*
  - $f(\vec{x})$ *is continuously differentiable, and* $\nabla f(\vec{x})$ *is Lipschitz continuous*
  - *Step size* μ *is such that:*
    - $\sum_{k=1}^{\infty} \mu_m = \infty; \sum_{k=1}^{\infty} (\mu_m)^2 < \infty$ *(The second term is not needed if there is no noise; recollect in steepest descent we used a constant; In machine learning, we will use a similar 'learning' rate, but never set it to a constant)*
  - For some scalars A and B, if:
    - $\mathbb{E}[\omega_m[i] | \mathcal{F}_m] = 0 \ \forall i;$
    - $\mathbb{E}\left[||\vec{\omega}_m||^2 \middle| \mathcal{F}_m \right] \leq A + B||\nabla f(\vec{x}_{\mathrm{m}})||^2$; $||\vec{\omega}_m||$ is the L2 norm (Euclidean distance)
    - $||\vec{\omega}_m|| = \sqrt{(\omega_m[1])^2 + (\omega_m[2])^2 + \cdots + (\omega_m[k])^2}$

Define Filtration $\mathcal{F}^m$: "History" of the algorithm up to and including m^{th} iteration

$\mathcal{F}^m = \{\vec{x}_0, \vec{x}_1, \ldots, \vec{x}_m, \vec{D}_0, \vec{D}_1, \ldots, \vec{D}_m, \mu_0, \mu_1, \ldots, \mu_m\}$; D: approximate estimations of derivatives

FDSA algorithm

- With probability 1, the sequence $\{f(\vec{x}_{\mathrm{m}})\}_{m=1}^{\infty}$ converges and $\lim_{m\to\infty} \nabla f(\vec{x}_{\mathrm{m}}) \to 0$, if
  - $f(\vec{x}) \geq 0$ *everywhere.*
  - $f(\vec{x})$ *is continuously differentiable, and* $\nabla f(\vec{x})$ *is Lipschitz continuous*
  - *Step size μ, such that:*
    - $\sum_{k=1}^{\infty} \mu_m = \infty; \sum_{k=1}^{\infty} (\mu_m)^2 < \infty$
  - For some scalars A and B, if
    - $\mathbb{E}[\omega_m[i] | \mathcal{F}_m] = 0 \ \forall i;$
    - $\mathbb{E}\left[||\vec{\omega}_m||^2 \middle| \mathcal{F}_m \right] \leq A + B||\nabla f(\vec{x}_{\mathrm{m}})||^2$

Make sure function is bounded, x is bounded

Pick step size that meets these conditions;
e.g., log(m)/m; A/(B+m); A=5, B=10
(In machine learning you will notice that a constant value is never used and the reason is it does not satisfy the second condition)

By central limit theorem, irrespective of distribution of random variables, estimation error through random sampling is Normal with 0 mean and finite variance ($\sim\mathcal{N}(0, \sigma^2)$) when sample is large enough)

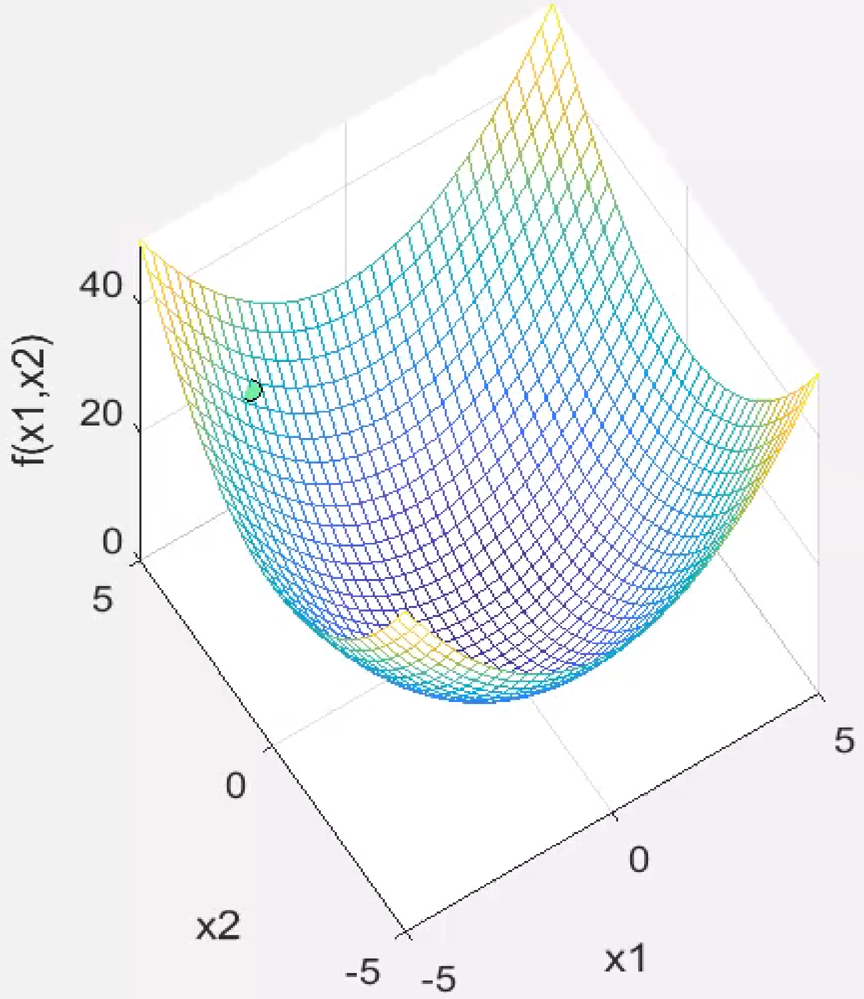
Define Filtration $\mathcal{F}^m$: "History" of the algorithm up to and including m^{th} iteration

$\mathcal{F}^m = \{\vec{x}_0, \vec{x}_1, \ldots, \vec{x}_m, \vec{D}_0, \vec{D}_1, \ldots, \vec{D}_m, \mu_0, \mu_1, \ldots, \mu_m\}$

D: approximate estimations of derivatives

SPSA Convergence?

- Black: actual gradient direction
- Red: SPSA approximation



Convergence conditions for stochastic approximations of GD (with noise (ω))

- With probability 1, the sequence $\{f(\vec{x}_\mathrm{m})\}_{m=1}^{\infty}$ converges and $\lim_{m\to\infty} \nabla f(\vec{x}_\mathrm{m}) \to 0$, if
  - *$f(\vec{x}) \geq 0$ everywhere.*
  - *$f(\vec{x})$ is continuously differentiable, and $\nabla f(\vec{x})$ is Lipschitz continuous*
  - *Step size μ is such that:*
    - $\sum_{k=1}^{\infty} \mu_m = \infty; \sum_{k=1}^{\infty} (\mu_m)^2 < \infty$ *(The second term is not needed if there is no noise; recollect in steepest descent we used a constant; In machine learning, we will use a similar 'learning' rate, but never set it to a constant)*
  - For some scalars A and B, if:
    - $\mathbb{E}[\omega_m[i] | \mathcal{F}_m] = 0 \ \forall i;$
    - $\mathbb{E}\left[\left|\left|\vec{\omega}_m\right|\right|^2 \middle| \mathcal{F}_m \right] \leq A + B\left|\left|\nabla f(\vec{x}_\mathrm{m})\right|\right|^2$; $\left|\left|\vec{\omega}_m\right|\right|$ is the L2 norm (Euclidean distance)
    - $\left|\left|\vec{\omega}_m\right|\right| = \sqrt{(\omega_m[1])^2 + (\omega_m[2])^2 + \cdots + (\omega_m[k])^2}$

Define Filtration $\mathcal{F}^m$: "History" of the algorithm up to and including m^{th} iteration

$\mathcal{F}^m = \{\vec{x}_0, \vec{x}_1, \ldots, \vec{x}_m, \vec{D}_0, \vec{D}_1, \ldots, \vec{D}_m, \mu_0, \mu_1, \ldots, \mu_m\}$; D: approximate estimations of derivatives

SPSA algorithm

- With probability 1, the sequence $\{f(\vec{x}_{\mathrm{m}})\}_{m=1}^{\infty}$ converges and $\lim_{m\to\infty} \nabla f(\vec{x}_{\mathrm{m}}) \to 0$, if
  - $f(\vec{x}) \geq 0$ *everywhere.*
  - $f(\vec{x})$ *is continuously differentiable, and* $\nabla f(\vec{x})$ *is Lipschitz continuous*
  - *Step size* μ*, such that:*
    - $\sum_{k=1}^{\infty} \mu_m = \infty; \sum_{k=1}^{\infty} (\mu_m)^2 < \infty$
  - For some scalars A and B, if
    - $\mathbb{E}[\omega_m[i]\,|\mathcal{F}_m] = 0\ \forall i;$
    - $\mathbb{E}\left[\left|\left|\vec{\omega}_m\right|\right|^2\middle|\mathcal{F}_m\right] \leq A + B\left|\left|\nabla f(\vec{x}_{\mathrm{m}})\right|\right|^2$

Make sure function is bounded, x is bounded

Pick step size that meets these conditions;
e.g., log(m)/m; A/(B+m); A=5, B=10
(In machine learning you will notice that a constant value is never used and the reason is it does not satisfy the second condition)

Next slide

Define Filtration $\mathcal{F}^m$: "History" of the algorithm up to and including m^{th} iteration

$\mathcal{F}^m = \{\vec{x}_0, \vec{x}_1, \ldots., \vec{x}_m, \vec{D}_0, \vec{D}_1, \ldots, \vec{D}_m, \mu_0, \mu_1, \ldots, \mu_m\}$

D: approximate estimations of derivatives

Start with Taylor's expansion for $f: R^2 \to R$

$$f\big(x(1)+h(1), x(2)+h(2)\big) \approx f\big(x(1), x(2)\big) + h(1)\frac{\partial f(\vec{x})}{\partial x(1)} + h(2)\frac{\partial f(\vec{x})}{\partial x(2)} + \frac{1}{2!}\left[h(1)^2\frac{\partial^2 f}{\partial x^2(1)} + 2h(1)h(2)\frac{\partial^2 f}{\partial x(1)\partial x(2)} + h(2)^2\frac{\partial^2 f}{\partial x^2(2)}\right]$$

$$f(x(1)-h(1), x(2)-h(2)) \approx f(x(1), x(2)) - h(1)\frac{\partial f}{\partial x(1)} - h(2)\frac{\partial f}{\partial x(2)} + \frac{1}{2!}h(1)^2\frac{\partial^2 f}{\partial x^2(1)} + \frac{1}{2!}2h(1)h(2)\frac{\partial^2 f}{\partial x(1)\partial x(2)} + \frac{1}{2!}h(2)^2\frac{\partial^2 f}{\partial x^2(2)}$$

$$f(\vec{x}+\vec{h}) - f(\vec{x}-\vec{h}) = 2h(1)\frac{\partial f}{\partial x(1)} + 2h(2)\frac{\partial f}{\partial x(2)}$$

$$\frac{\partial f}{\partial x(1)} = \frac{\left(f(\vec{x}+\vec{h}) - f(\vec{x}-\vec{h})\right)}{2h(1)} + \frac{2h(2)}{2h(1)}\frac{\partial f}{\partial x(2)}$$ (SPSA estimation for derivative excludes the second expression, and thus, represents the estimation error)

$$\frac{\partial f}{\partial x(1)} = \frac{\left(f(\vec{x}+\vec{h}) - f(\vec{x}-\vec{h})\right)}{2h(1)} + error$$

General expression for $f: R^k \to R$

$$\frac{f(\vec{x}+\vec{h}) - f(\vec{k}-\vec{h})}{2h(i)} = \frac{\partial f}{\partial x(i)} + \sum_{i\neq i, j=1}^{k}\frac{h_m(j)}{h_m(i)}\frac{\partial f}{\partial x(j)}$$; m indicates it is the error at m^{th} iteration of SPSA

$error = e = \sum_{i\neq i, j=1}^{k}\frac{h_m(j)}{h_m(i)}\frac{\partial f}{\partial x(j)}$; to show convergence prove that

$$\mathbb{E}[e_m[i] | \mathcal{F}_m] = 0 \ \forall i;$$

$$\mathbb{E}\left[\left|\left|\vec{e}_m\right|\right|^2 \Big| \mathcal{F}_m\right] \leq A + B\left|\left|\nabla f(\vec{x}_{\mathrm{m}})\right|\right|^2$$

Prove: $\mathbb{E}[e_m[i]|\mathcal{F}_m] = 0\ \forall i; \mathbb{E}\left[||\vec{e}_m||^2 \middle| \mathcal{F}_m\right] \le A + B||\nabla f(\vec{x}_\mathrm{m})||^2$

- $E[e(i)|\mathcal{F}_m] = \sum_{\mathrm{j}\neq i,\mathrm{j}=1}^{k} \mathbb{E}\left[\frac{h_m(\mathrm{j})}{h_m(i)}\frac{\partial f}{\partial x(j)}|\mathcal{F}_m\right] = \sum \left\{(0.5\times -1 + 0.5\times 1)\mathbb{E}\left[\frac{\partial f}{\partial x(\mathrm{j})}|\mathcal{F}_m\right]\right\} = 0$
  - (satisfies first condition $\mathbb{E}[e_m[i]|\mathcal{F}_m] = 0\ \forall i;$)
- $||\mathrm{e}||^2 = \mathrm{e}(1)^2 + \mathrm{e}(2)^2 + \cdots. + \mathrm{e}(\mathrm{k})^2 = \left[\sum_{\mathrm{j}\neq 1,\mathrm{j}=1}^{k}\left[\frac{h_m(\mathrm{j})}{h_m(1)}\frac{\partial f}{\partial x(j)}\right]\right]^2 + \cdots + \left[\sum_{\mathrm{j}\neq \mathrm{k},\mathrm{j}=1}^{k}\left[\frac{h_m(\mathrm{j})}{h_m(k)}\frac{\partial f}{\partial x(j)}\right]\right]^2$
- $= \sum_{\mathrm{j}\neq 1,\mathrm{j}=1}^{k}\left[\frac{\partial f}{\partial x(j)}\right]^2 \frac{h_m(\mathrm{j})^2}{h_m(1)^2} + \cdots + \sum_{\mathrm{j}\neq k,\mathrm{j}=1}^{k}\left[\frac{\partial f}{\partial x(j)}\right]^2 \frac{h_m(\mathrm{j})^2}{h_m(k)^2} + A;$
- $\frac{h_m(\mathrm{j})^2}{h_m(1)^2}$=1; A= sum of product of derivates, which we assume are bounded as derivatives are bounded (look at Lipschitz continuity)
- $\Rightarrow ||\mathrm{e}||^2 = (k-1)||\nabla f(\vec{x}_\mathrm{m})||^2 + A$ (satisfies second condition $\mathbb{E}\left[||\vec{e}_m||^2 \middle| \mathcal{F}_m\right] \le A + B||\nabla f(\vec{x}_\mathrm{m})||^2$)

$h(i)$ and $h(j)$ are random selections from [-1,1];
$\Rightarrow \mathbb{E}\left[\frac{h_m(\mathrm{j})}{h_m(i)}\right] =?$

References

- Spall, J. C., “Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation,” IEEE Trans. Autom. Control 37, 332–341 (1992).
- "An Overview of the Simultaneous Perturbation Method for Efficient Optimization,“ Johns Hopkins APL Technical Digest, vol. 19(4), pp. 482–492.
- Simulation based optimization, by Gosavi
  - Chapters (Parametric Optimization: Stochastic Gradients and Adaptive Search) and
  - Convergence in Chapter ‘Convergence Analysis of Parametric Optimization Methods’