# II. Descent Methods: Descent for quadratic norm

1. Problem: Min $f(x)$
2. For each iteration, we try the steepest descent in terms of a given norm.

$$Min_{\Delta x} \ \nabla f(x)^T \Delta x$$
$$\text{s.t. } \|\Delta x\|_P \le 1 \Rightarrow \|\Delta x\|_P - 1 \le 0$$

$\|\Delta x\|_P = (\Delta x^T P \Delta x)^{1/2}, P \in S^n_{++}$

Lagrangian $L(\Delta x, \lambda) = \nabla f(x)^T \Delta x + \lambda(\|\Delta x\|_P - 1), \lambda \ge 0$

We can derive: $\Delta x_{nsd} = -\left(\nabla f(x)^T P^{-1} \nabla f(x)\right)^{-1/2} P^{-1} \nabla f(x)$

Or $\Delta x_{sd} = -P^{-1} \nabla f(x)$



# II. Descent Methods: Descent for quadratic norm

The coordinate change has effects on the descent direction.

Example: min $f(x) = \frac{1}{2} x^T P x + q^T x, P \in S^n_{++}$

Affine transform: $\bar{x} = P^{1/2} x$

$$f(\bar{x}) = \frac{1}{2} \bar{x}^T \bar{x} + q^T P^{-1/2} \bar{x}$$

$$\nabla_{\bar{x}} f(\bar{x}) = \bar{x} + P^{-1/2} q$$

$$\bar{x} = -P^{-1/2} q$$

$$\text{Or } x = -P^{-1} q$$

$$\nabla_x f(x) = Px + q = 0$$

$$x = -P^{-1} q$$

# II. Descent Methods: Example

$[x_1 \;\; x_2]\begin{bmatrix} 1 & 0 \\ 0 & \gamma \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

Problem: $\min f(x) = \frac{1}{2}(x_1^2 + \gamma x_2^2) \quad \gamma > 0$

$$x^o = (\gamma, 1), f(x^0) = \frac{\gamma(\gamma+1)}{2}, \nabla f(x^0) = (\gamma, \gamma)$$

Thus, $x^1 = (\gamma, 1) - t(\gamma, \gamma) = (\gamma(1-t), 1 - t\gamma)$

and $\nabla f(x^1) = (\gamma(1-t), \gamma(1-t\gamma))$

1. To opt $f(x^1)$ with respect to variable $t$,

we have $f(x^1) = \frac{1}{2}(\gamma^2(1-t)^2 + \gamma(1-t\gamma)^2)$

$$\frac{\partial f(x^1)}{\partial t} = \gamma^2(1-t) + \gamma(1-t\gamma)\gamma = 0$$

Thus, $t = \frac{2\gamma^2}{\gamma^2+\gamma^3} = \frac{2}{1+\gamma}$, and $x^1 = \left(\frac{\gamma(\gamma-1)}{1+\gamma}, \frac{1-\gamma}{1+\gamma}\right) = \left(\frac{10\times 9}{11}, -\frac{9}{11}\right)$

2. We repeat the process to step $k$, $x^k = (\gamma\left(\frac{\gamma-1}{\gamma+1}\right)^k, \left(\frac{1-\gamma}{1+\gamma}\right)^k)$

3. Equal potential plot

$$f(x^k) = \frac{\gamma(\gamma+1)}{2}\left(\frac{\gamma-1}{\gamma+1}\right)^{2k} = \left(\frac{\gamma-1}{\gamma+1}\right)^{2k} f(x^o) = \left(\frac{1-m/M}{1+m/M}\right)^{2k} f(x^o)$$



# II. Descent Methods: Descent for various norms

1. Problem: Min $f(x)$
2. For each iteration, we try the steepest descent in terms of a given norm.

$$\text{Min } \nabla f(x)^T \Delta x$$
$$\text{s.t. } \|\Delta x\| \leq 1$$

3. We show the step of
   i. Quadratic norm
   ii. L1 norm



9WA

$$L(\Delta x,\lambda) = \nabla f(x)^T \Delta x + \lambda (\Delta x^T P \Delta x)^{1/2} - \lambda$$

$\frac{1}{2}\ 2$

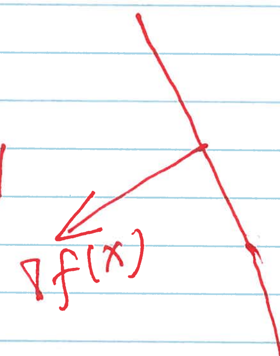
$$\nabla_{\Delta x} L(\Delta x,\lambda) = \nabla f(x) + \lambda P \Delta x (\Delta x^T P \Delta x)^{-1/2} = 0$$

Thus, we have

$$\Delta x = -\frac{(\Delta x^T P \Delta x)^{1/2}}{\lambda} P^{-1} \nabla f(x)$$

Utilize the constraint

$$\|\Delta x\|_P = 1 \qquad (\Delta x^T P \Delta x)^{1/2} = 1$$




We have

$$\Delta x = -[\nabla f(x)^T P^{-1} \nabla f(x)]^{-1/2} P^{-1} \nabla f(x)$$

Remark ① $[\nabla f(x)^T P^{-1} \nabla f(x)]^{-1/2}$ normalize the magnitude

② The descent direction changes by $P^{-1}$.

$$f(x) = \frac{1}{2} x^T P x + q^T x$$

① $\nabla_x f(x) = Px + q$

② Let $\bar{x} = P^{1/2} x$

$$f(\bar{x}) = \frac{1}{2} \bar{x}^T \bar{x} + q^T P^{-1/2} \bar{x}$$

$$\nabla_{\bar{x}} f(\bar{x}) = \bar{x} + P^{-1/2} q$$

$$\text{or} \quad P^{+1/2} x + P^{-1/2} q$$

Remark ① The gradient changes with the transform

② The quadratic analytic solution remains the same.

# II. Descent Methods: Descent for L1 norm

1. Problem: Min $f(x)$
2. For each iteration, we try the steepest descent in terms of a given norm.

$$\text{Min } \nabla f(x)^T \Delta x < 0$$
$$\text{s.t. } \|\Delta x\|_1 \leq 1$$

Lagrangian $L(\Delta x, \lambda) = \nabla f(x)^T \Delta x + \lambda\,(\|\Delta x\|_1 - 1),\ \lambda \geq 0$

We can derive: $\Delta x_{nsd} = -sign\left(\frac{\partial f(x)}{\partial x_i}\right) e_i$,

where $i$ is the index for which $\|\nabla f(x)\|_\infty = |\nabla f(x)_i|$

Or $\Delta x_{sd} = -\frac{\partial f(x)}{\partial x_i} e_i$



# Gradient descent method: Convergence analysis

$$\tilde{f}(t) \equiv f(x - t\nabla f(x)) \leq f(x) - t\|\nabla f(x)\|_2^2 + \frac{Mt^2}{2}\|\nabla f(x)\|_2^2$$

$$\tilde{f}(t_{exact}) \leq \tilde{f}\left(t = \frac{1}{M}\right) \leq f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2 \quad (\min_t f(x)\nabla f(x))$$

A. $\tilde{f}(t_{exact}) - p^* \leq f(x) - p^* - \frac{1}{2M}\|\nabla f(x)\|_2^2$

B. $\frac{1}{2M}\|\nabla f(x)\|_2^2 \geq \frac{m}{M}(f(x) - p^*)$ since $\frac{\|\nabla f(x)\|_2^2}{2m} \geq f(x) - p^*$

C. From B, we have

$$f(x) - p^* - \frac{1}{2M}\|\nabla f(x)\|_2^2 \leq f(x) - p^* - \frac{m}{M}(f(x) - p^*)$$
$$= (f(x) - p^*)(1 - \frac{m}{M})$$

D. We can conclude from A & C

$$f(x^{k+1}) - p^* \leq \left(1 - \frac{m}{M}\right)(f(x^k) - p^*) \leq \left(1 - \frac{m}{M}\right)^k (f(x^o) - p^*)$$

To achieve $f(x^*) - p^* \leq \epsilon$,

we need $\frac{\log((f(x^o) - p^*)/\epsilon)}{\log(1/c)}$ $steps$, where $c = 1 - \frac{m}{M} < 1$,

## Gradient descent method : Convergence analysis

$\log(1/c) = -\log(1 - m/M) \approx m/M$ for large $M/m$

Remark: when $M/m > 100$

the method can be very slow.



## Newton Step

Use the approximation of 2nd order Taylor's Exp.

$$f(x+v) \approx f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v$$

$[-\nabla^2 f(x)^{-1} \nabla f(x)]^T \nabla^2 f(x)$

We would like to derive

$\nabla_v f(x+v) = 0 \rightarrow \nabla f(x) + \nabla^2 f(x) v = 0$

$[-\nabla^2 f(x)^{-1} \nabla f(x$

Thus, we have $v = -\nabla^2 f(x)^{-1} \nabla f(x)$

$$\begin{aligned} f(x+v) &= f(x) + (-1)\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) + \\ &\quad \underline{\tfrac{1}{2} \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)} \\ &= f(x) - \tfrac{1}{2} \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) \end{aligned}$$

Input $x \in dom\, f,\ \epsilon > 0$

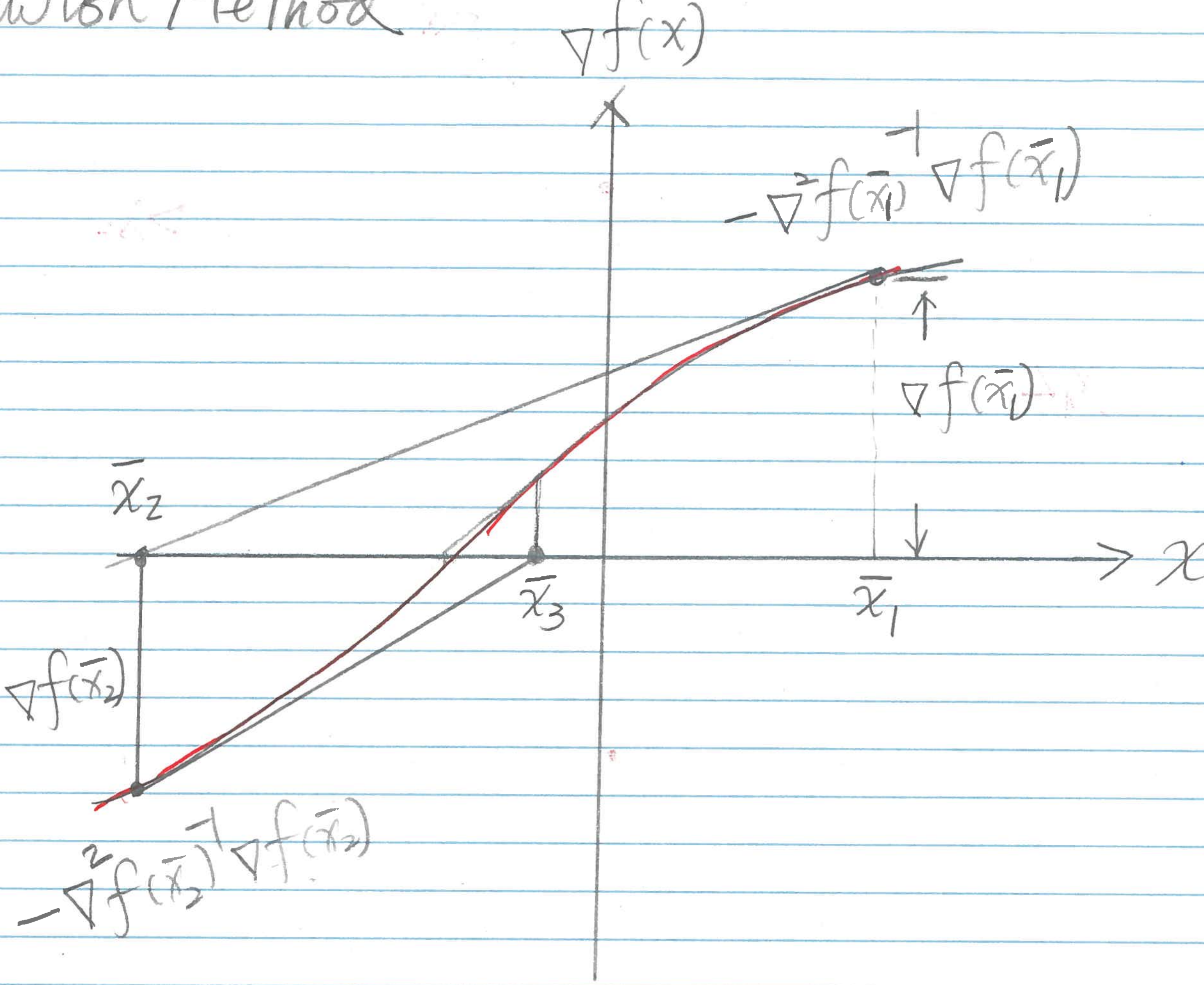
Repeat: 1. $\Delta x_{nt} := -\nabla^2 f(x)^{-1} \nabla f(x), \lambda^2(x) = \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)$

2. $Quit\ if\ \lambda^2/2 \le \epsilon$

3. $Line\ Search\ t$

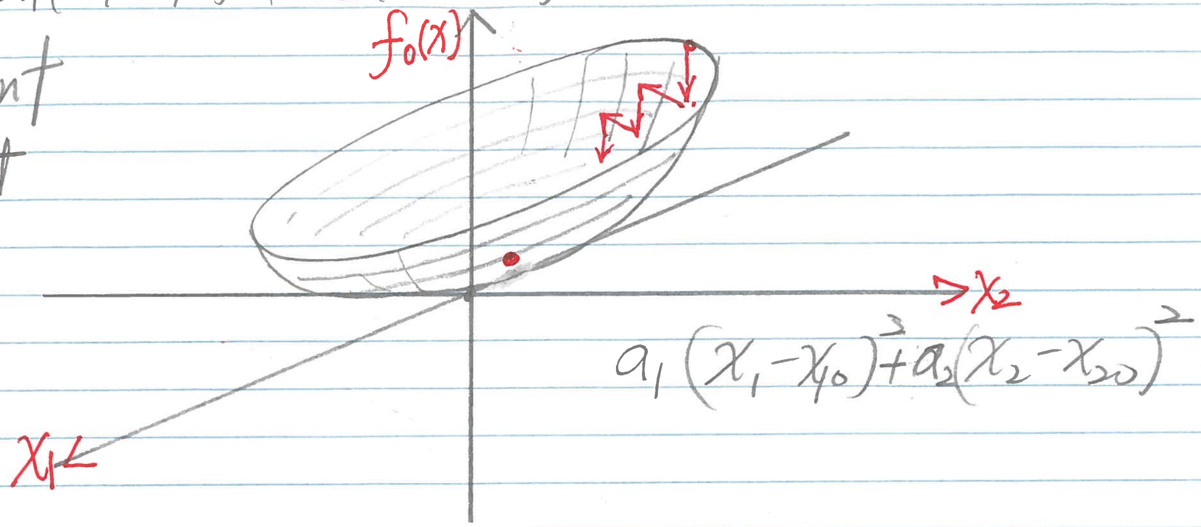
4. $x := x + t\Delta x_{nt}$

# Newton Method

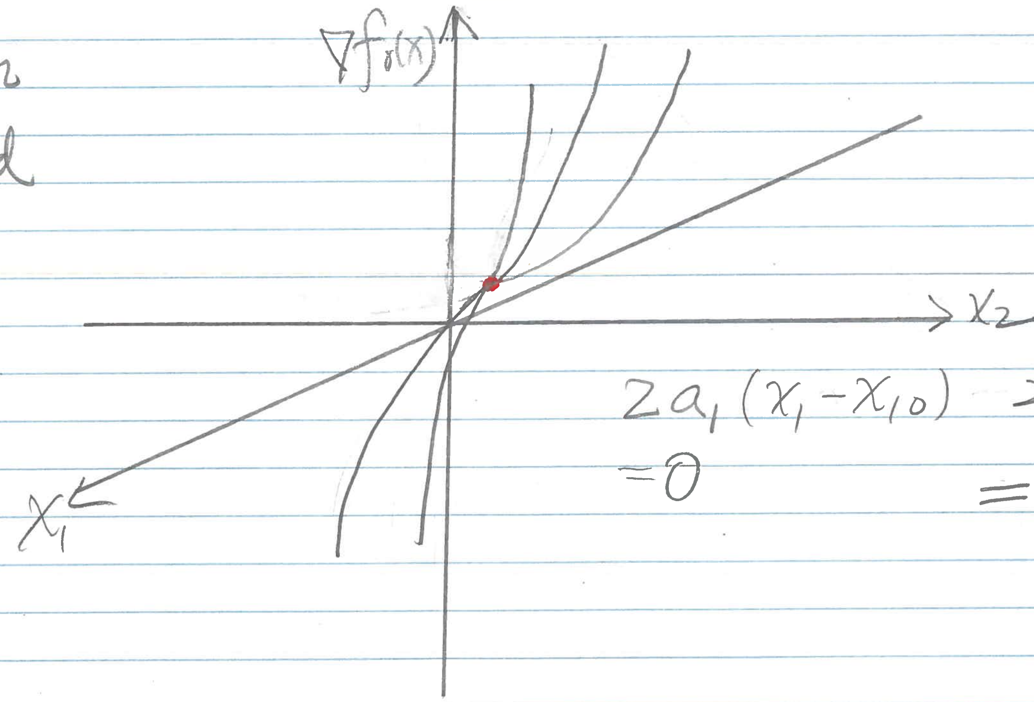


$$\min \; a_1(x_1-x_{10})^2 + a_2(x_2-x_{20})^2$$

Gradient descent




Newton Method




$2a_1(x_1-x_{10}) = 0 \qquad 2a_2(x_2-x_{20}) = 0.$

# Newton Method : Convergence analysis

Assumptions: $S = \{x \in dom\, f \mid f(x) \le f(x_o)\}$

$f$ strongly convex on S with constant m, s.t. $\nabla^2 f(x) \ge mI, \forall x \in S$

$\nabla^2 f$ is Lipschitz continuous on S with constant $L$, i.e.

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \le L\|x-y\|_2$$

Outlines: $\exists \eta \in (0, m^2/L)$, two cases.

1. Damped Newton Phase: $(t < 1)$

$\|\nabla f(x)\|_2 \ge \eta$ then $f(x^{k+1}) - f(x^k) \le -\alpha\beta\eta^2 m/M^2$

2. Pure Newton Phase (Quadratically Convergent Stage): $(t = 1)$

$\|\nabla f(x)\|_2 < \eta$ then

$$\frac{L}{m^2}\left\|\nabla f(x^{k+1})\right\|_2 \le \left(\frac{L}{2m^2}\left\|\nabla f(x^k)\right\|_2\right)^2$$

$$\le \left(\frac{L}{2m^2}\left\|\nabla f(x^l)\right\|_2\right)^{2^{k+1-l}} \le \left(\frac{1}{2}\right)^{2^{k+1-l}} \qquad k+1 \ge l$$

$(1-m/M)^k$ ① $0.9,\ 0.9^2=0.81,\ 0.9^3=0.73,\ 0.9^4=0.63,\ 0.9^5=0.54,\ 0.9^6=0.48$



② $0.9,\ 0.9^2=0.81,\ (0.81)^2=0.64,\ (0.64)^2=0.36,\ (0.36)^2=0.16$

$(0.16)^2=0.03,\ (0.03)^2=0.0009,\ (0.0009)^2 \approx 0.000000\ldots$

$10^{-3}$ $10^{-6}$

# Newton Method: Affine Invariant

Problem: $\min f(x)$

Theorem: Newton's step is invariant to affine transform.

Proof: Let $x = Ty, T \in R^{nn}, f(x) = f(Ty) = \bar{f}(y)$

For the $x$ coordinate system, we have.

$\Delta x_{nt} = -\nabla^2 f(x)^{-1}\nabla f(x)$

Therefore, we have the invariant results

$x + \Delta x_{nt} = T(y + \Delta y_{nt})$.

For the $y$ coordinate system, we have.

1. $\nabla_y \bar{f}(y) = T^T \nabla_x f(Ty)$,

   $\nabla_y^2 \bar{f}(y) = T^T \nabla^2 f(Ty) T$

2. The Newton step at $y$,

$\Delta y_{nt}$

$= -\nabla_y^2 \bar{f}(y)^{-1}\nabla_y \bar{f}(y)$

$= -(T^T\nabla^2 f(x)T)^{-1}(T^T\nabla f(x))$

$= -T^{-1}\nabla^2 f(x)^{-1}\nabla f(x)$

$= T^{-1}\,\Delta x_{nt}$

# Summary

1. Gradient Descent Method: (minimization solution)
    1. Vector operations per iteration
    2. Linear convergence rate
2. Newton's Method: (equality solution)
    1. Matrix operations per iteration
    2. Quadratic convergence rate (near the solution)
3. Gradient Descent Method Variations:
    1. Conjugate gradient method
    2. Nesterov gradient descent method
    3. Quasi-Newton method