Hunting Inverse Hessian Matrices

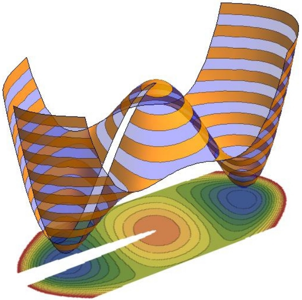
Robert Gower
and Jacek Gondzio

Irish Applied Mathematics Research Students' Meeting 2014, Galway.

December 11, 2014

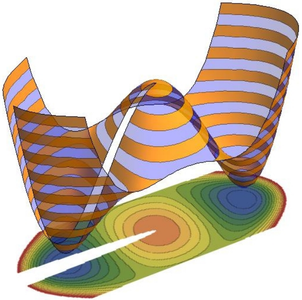
Nonlinear optimization is where I hunt

$$\min_x f(x)$$
with $g(x) \leq 0$.



"Nothing in the world takes place without optimization, and there is no doubt that all aspects of the world that have a rational basis can be explained by optimization methods", Leonhard Euler (1744).

Nonlinear optimization is where I hunt



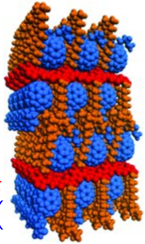
$$\min_x f(x)$$
with $g(x) \leq 0$.

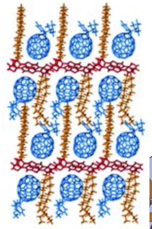
"Nothing in the world takes place without optimization, and there is no doubt that all aspects of the world that have a rational basis can be explained by optimization methods", Leonhard Euler (1744).

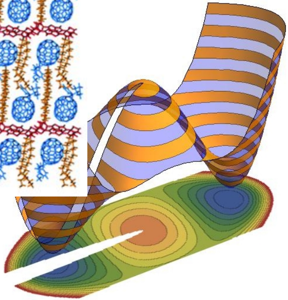
Nonlinear optimization is where I hunt

$\min_x$

with $g($

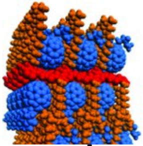


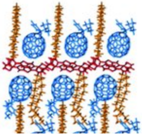


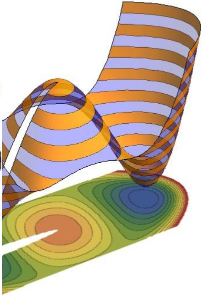


"Nothing in the world takes place without optimization, and there is no doubt that all aspects of the world that have a rational basis can be explained by optimization methods", Leonhard Euler (1744).

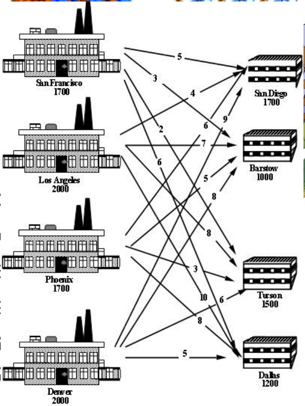
Nonlinear optimization is where I hunt





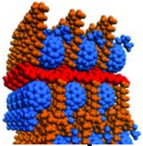


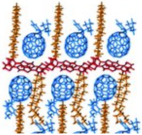
$\color{red}{\min_x}$
with $\color{blue}{g(}$

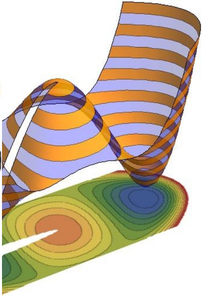


"Nothing in the wo
without optimizati
doubt that all aspe
that have a rationa
explained by optim
Leonhard Euler (17

Nonlinear optimization is where I hunt

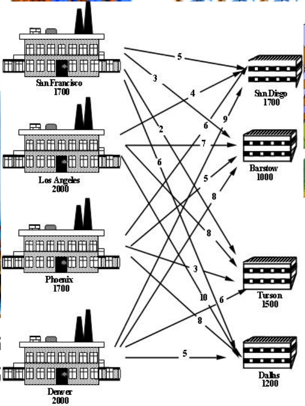






$\min_x$

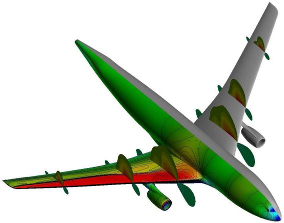
with $g($

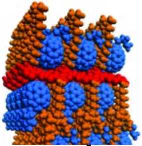


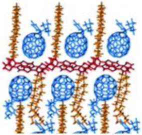
explained by optim

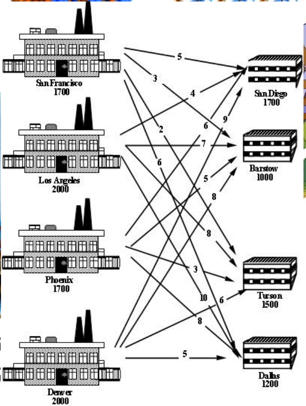
Leonhard Euler (17

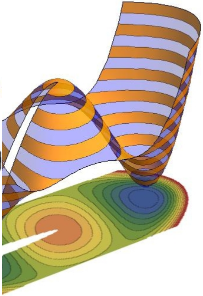
Nonlinear optimization is where I hunt







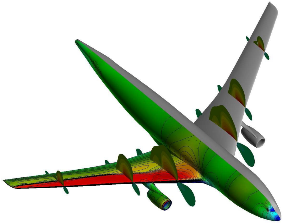


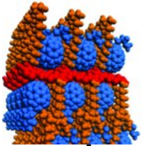


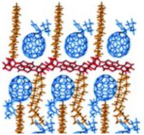
explained by optim

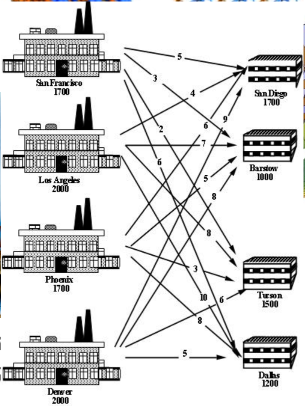
Leonhard Euler (17

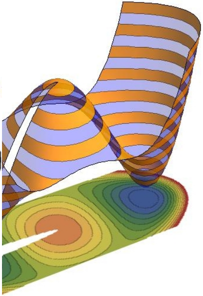
Nonlinear optimization is where I hunt







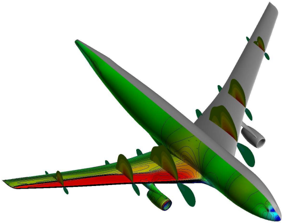


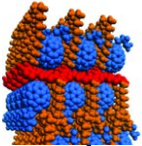


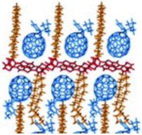
explained by optim

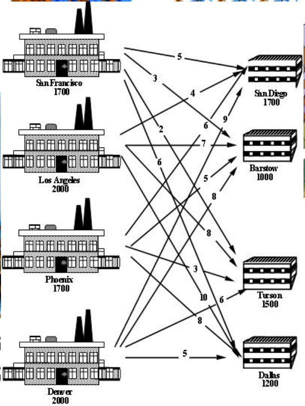
Leonhard Euler (17

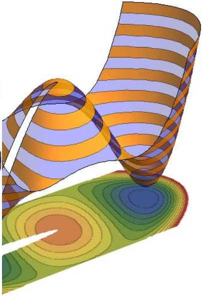
Nonlinear optimization is where I hunt







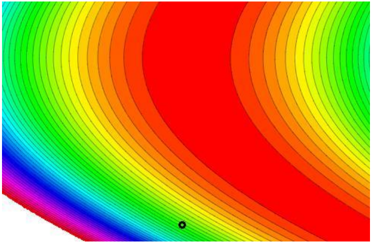


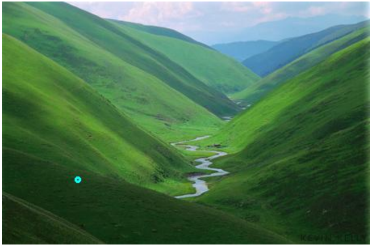


explained by optim

Leonhard Euler (17

$\min_x f(x) = 100(x - y^2)^2 + (y-1)^2$



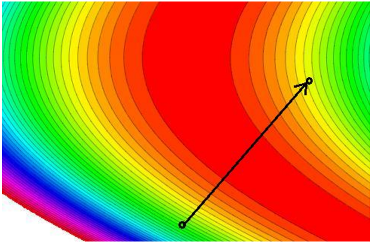


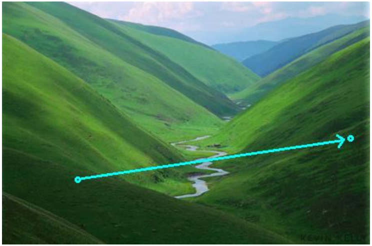
Difficult global problem. Use a local information.

Input: $x_0 \in \mathbb{R}^n$
for $k = 0, 1, 2, \ldots$ **do**
  $x_{k+1} = x_k - \nabla f(x_k)$
end

Zigzags 4'000 iterations!

$$\min_x f(x) = 100(x - y^2)^2 + (y - 1)^2$$



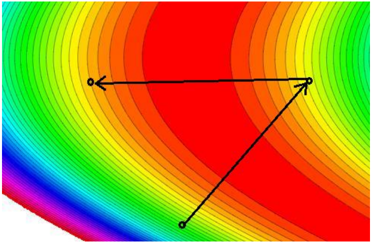


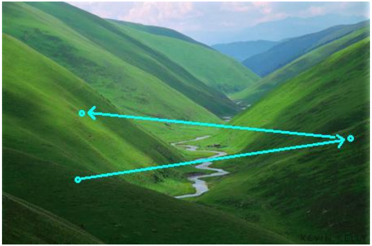
Difficult global problem. Use a local information.

Input: $x_0 \in \mathbb{R}^n$
for $k = 0, 1, 2, \dots$ **do**
| $x_{k+1} = x_k - \nabla f(x_k)$
end

Zigzags 4'000 iterations!

$\min_x f(x) = 100(x - y^2)^2 + (y - 1)^2$





Difficult global problem. Use a local information.

Input: $x_0 \in \mathbb{R}^n$

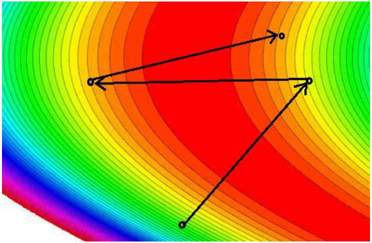
for $k = 0, 1, 2, \dots$ **do**

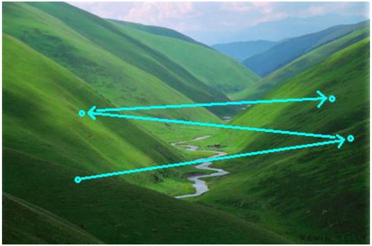
$\quad x_{k+1} = x_k - \nabla f(x_k)$

end

Zigzags 4'000 iterations!

$\min_x f(x) = 100(x - y^2)^2 + (y - 1)^2$





Difficult global problem. Use a local information.

Input: $x_0 \in \mathbb{R}^n$
for $k = 0, 1, 2, \ldots$ **do**
  $x_{k+1} = x_k - \nabla f(x_k)$
end

Zigzags 4'000 iterations!

Using Second-order information

Search for stationary point

$$\nabla f(x) = 0, \quad \text{Fermat 1646}$$

linearize around x_k

$$\nabla f(x_k + d) \approx \nabla^2 f(x_k) d + \nabla f(x_k)$$

Using Second-order information

Search for stationary point

$$\nabla f(x) = 0, \qquad \text{Fermat 1646}$$

linearize around x_k

$$\nabla f(x_k + d) \approx \nabla^2 f(x_k) d + \nabla f(x_k) = 0$$

Using Second-order information

Search for stationary point

$$\nabla f(x) = 0, \qquad \text{Fermat 1646}$$

linearize around x_k

$$\nabla f(x_k + d) \approx \nabla^2 f(x_k) d + \nabla f(x_k) = 0$$

Newton's Method

Input: $x_0 \in \mathbb{R}^n$

for $k = 0, 1, 2, \ldots$ **do**

  Solve $\nabla^2 f(x_k) d_k = -\nabla f(x_k)$

  $x_{k+1} = x_k + d_k$

end

Using Second-order information

Search for stationary point

$$\nabla f(x) = 0, \quad \text{Fermat 1646}$$

linearize around x_k

$$\nabla f(x_k + d) \approx \nabla^2 f(x_k) d + \nabla f(x_k) = 0$$

Newton's Method

Input: $x_0 \in \mathbb{R}^n$

for $k = 0, 1, 2, \ldots$ **do**

  Solve $\boxed{\nabla^2 f(x_k) d_k = -\nabla f(x_k)}$

  $x_{k+1} = x_k + d_k$

end

Solve a linear system

Using Second-order information

Search for stationary point

$$\nabla f(x) = 0, \quad \text{Fermat 1646}$$

linearize around x_k

$$\nabla f(x_k + d) \approx \nabla^2 f(x_k) d + \nabla f(x_k) = 0$$

Newton's Method

Input: $x_0 \in \mathbb{R}^n$

for $k = 0, 1, 2, \ldots$ **do**

  Solve $\boxed{\nabla^2 f(x_k) d_k = -\nabla f(x_k)}$

  $x_{k+1} = x_k + d_k$

end

Solve a linear system

Solving one Newton system

Proxy solve $\nabla^2 f(x_k) d_k = -\nabla f(x_k)$

$$d_k = \min_{d \in \mathcal{S}_k} \|\nabla^2 f(x_k) d + \nabla f(x_k)\|$$

where $\mathcal{S}_k \subset \mathbb{R}^n$ is a subspace.
Requires calculating $\nabla^2 f_k \mathcal{S}_k$

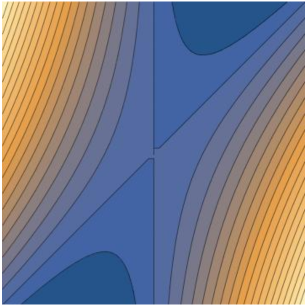


Figure: Contour $d^T \nabla^2 f_k d$

Problem: Solving linears system expensive. What can be done?

Solving one Newton system

Proxy solve $\nabla^2 f(x_k) d_k = -\nabla f(x_k)$

$$d_k = \min_{d \in \mathcal{S}_k} \|\nabla^2 f(x_k) d + \nabla f(x_k)\|$$

where $\mathcal{S}_k \subset \mathbb{R}^n$ is a subspace.
Requires calculating $\nabla^2 f_k \mathcal{S}_k$

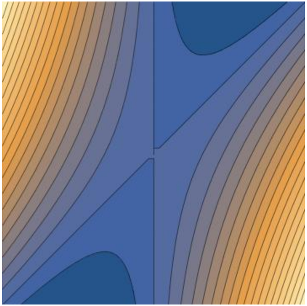


Figure: Contour $d^T \nabla^2 f_k d$

Problem: Solving linears system expensive. What can be done?
Another interpretation: Stationary points of local quadratic

$$f(x_k + d) \approx f(x_k) + \langle \nabla f(x_k), d \rangle + \frac{1}{2} d^T \nabla^2 f(x_k) d.$$

Solving one Newton system

Proxy solve $\nabla^2 f(x_k) d_k = -\nabla f(x_k)$

$$d_k = \min_{d \in \mathcal{S}_k} \|\nabla^2 f(x_k) d + \nabla f(x_k)\|$$

where $\mathcal{S}_k \subset \mathbb{R}^n$ is a subspace.
Requires calculating $\nabla^2 f_k \mathcal{S}_k$

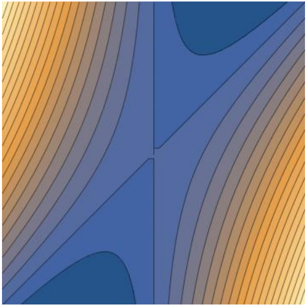


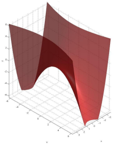
Figure: Contour $d^T \nabla^2 f_k d$

Problem: Solving linears system expensive. What can be done?
Another interpretation: Stationary points of local quadratic

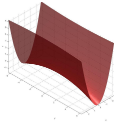
$$f(x_k + d) \approx f(x_k) + \langle \nabla f(x_k), d \rangle + \frac{1}{2} d^T \nabla^2 f(x_k) d.$$

Newton's Method

$$\boxed{\begin{aligned} \nabla^2 f(x_k) d_k &= -\nabla f(x_k). \\ x_{k+1} &= x_k + d_k \end{aligned}}$$



x_0



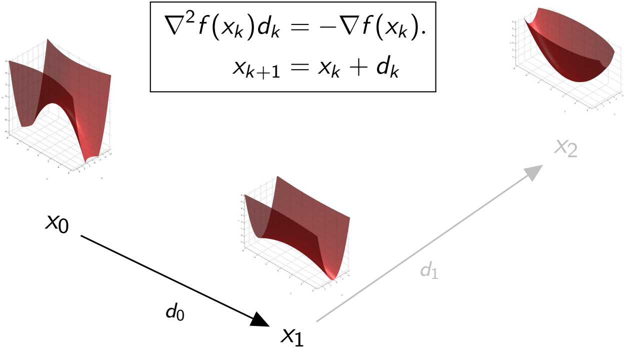
d_0

x_1

The Hessian "slowly" changes $\nabla^2 f(x_{k+1}) \approx \nabla^2 f(x_k)$.

Newton's Method

$$\boxed{\begin{aligned} \nabla^2 f(x_k) d_k &= -\nabla f(x_k). \\ x_{k+1} &= x_k + d_k \end{aligned}}$$

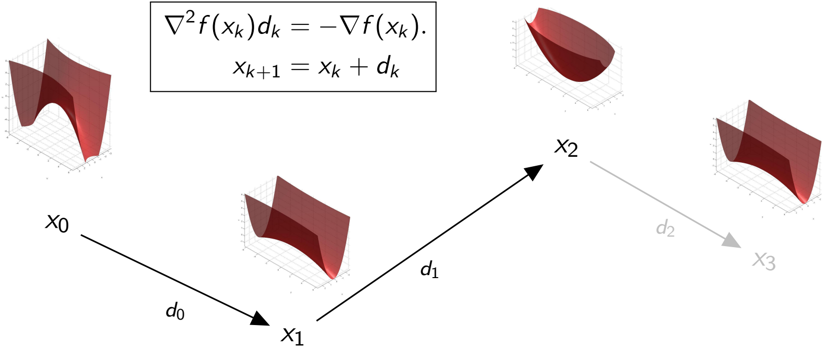


The Hessian "slowly" changes $\nabla^2 f(x_{k+1}) \approx \nabla^2 f(x_k)$.

Each Newton system is similar

Newton's Method

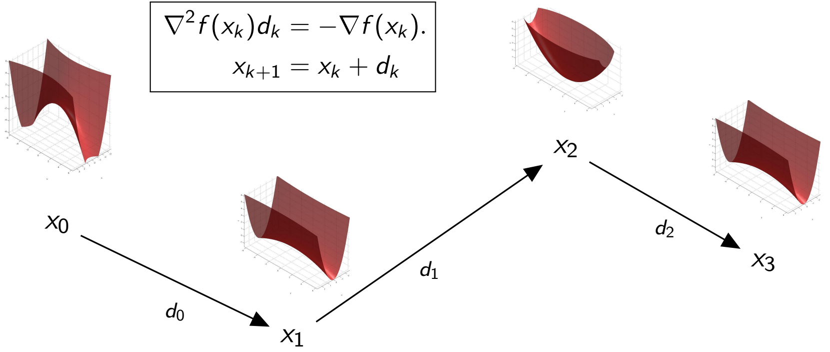
$$\nabla^2 f(x_k) d_k = -\nabla f(x_k).$$
$$x_{k+1} = x_k + d_k$$



The Hessian "slowly" changes $\nabla^2 f(x_{k+1}) \approx \nabla^2 f(x_k)$.
Each Newton system is similar
Solving each system individually is a waste.

Newton's Method

$$\nabla^2 f(x_k) d_k = -\nabla f(x_k).$$
$$x_{k+1} = x_k + d_k$$



The Hessian "slowly" changes $\nabla^2 f(x_{k+1}) \approx \nabla^2 f(x_k)$.
Each Newton system is similar
Solving each system individually is a waste.

Solving one Newton system **using the previous**

Proxy solve $\nabla^2 f(x_k) d_k = -\nabla f(x_k)$

$$d_k = \arg\min_{d \in \mathcal{S}_k} \|\nabla^2 f(x_k) d + \nabla f(x_k)\|$$

where $\mathcal{S}_k \subset \mathbb{R}^n$ is a subspace.
Requires calculating $\nabla^2 f_k \mathcal{S}_k$

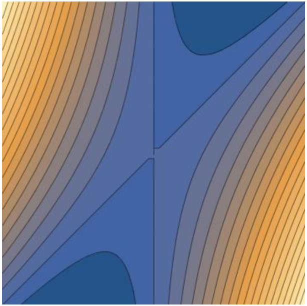


Figure: $d^T \nabla^2 f_k d$

Changing Coordinates $d = Py$ can help when

Solving one Newton system **using the previous**

Proxy solve $\nabla^2 f(x_k) d_k = -\nabla f(x_k)$

$$d_k = \arg \min_{Py \in \mathcal{S}_k} \|\nabla^2 f(x_k) P y + \nabla f(x_k)\|$$

where $\mathcal{S}_k \subset \mathbb{R}^n$ is a subspace.
Requires calculating $\nabla^2 f_k \mathcal{S}_k$

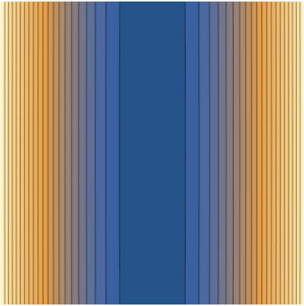


Figure: $y^T P^T \nabla^2 f_k P y$

Changing Coordinates $d = Py$ can help when
$P \approx \nabla^2 f_k^{-1}$ then $\nabla^2 f_k P \approx I$, easy.

Solving one Newton system **using the previous**

Proxy solve $\nabla^2 f(x_k) d_k = -\nabla f(x_k)$

$$d_k = \arg\min_{Py \in \mathcal{S}_k} \|\nabla^2 f(x_k) Py + \nabla f(x_k)\|$$

where $\mathcal{S}_k \subset \mathbb{R}^n$ is a subspace.
Requires calculating $\nabla^2 f_k \mathcal{S}_k$

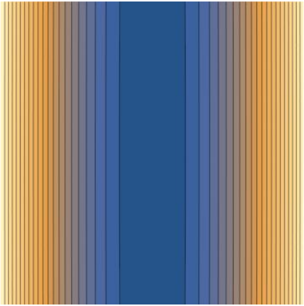


Figure: $y^T P^T \nabla^2 f_k P y$

Changing Coordinates $d = Py$ can help when $P \approx \nabla^2 f_k^{-1}$ then $\nabla^2 f_k P \approx I$, easy.

Objective: $P_{k-1} \approx \nabla^2 f(x_{k-1})^{-1}$ from available information to precondition $\|\nabla^2 f(x_k) P_{k-1} y + \nabla f(x_k)\|$.

Solving one Newton system **using the previous**

Proxy solve $\nabla^2 f(x_k) d_k = -\nabla f(x_k)$

$$d_k = \arg \min_{Py \in \mathcal{S}_k} \|\nabla^2 f(x_k) Py + \nabla f(x_k)\|$$

where $\mathcal{S}_k \subset \mathbb{R}^n$ is a subspace.
Requires calculating $\nabla^2 f_k \mathcal{S}_k$

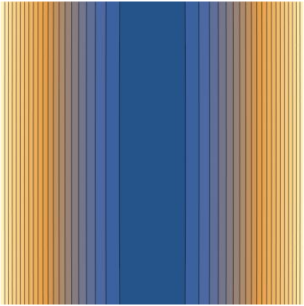


Figure: $y^T P^T \nabla^2 f_k P y$

Changing Coordinates $d = Py$ can help when $P \approx \nabla^2 f_k^{-1}$ then $\nabla^2 f_k P \approx I$, easy.

$\boxed{\text{Objective: } P_{k-1} \approx \nabla^2 f(x_{k-1})^{-1} \text{ from available information to precondition } \|\nabla^2 f(x_k) P_{k-1} y + \nabla f(x_k)\|.}$

Solving one Newton system **using the previous**

Proxy solve $\nabla^2 f(x_k) d_k = -\nabla f(x_k)$

$$d_k = \arg \min_{Py \in \mathcal{S}_k} \|\nabla^2 f(x_k) P y + \nabla f(x_k)\|$$

where $\mathcal{S}_k \subset \mathbb{R}^n$ is a subspace.
Requires calculating $\nabla^2 f_k \mathcal{S}_k$

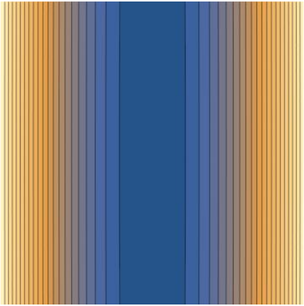


Figure: $y^T P^T \nabla^2 f_k P y$

Changing Coordinates $d = Py$ can help when
$P \approx \nabla^2 f_k^{-1}$ then $\nabla^2 f_k P \approx I$, easy.
Objective: $P_{k-1} \approx \nabla^2 f(x_{k-1})^{-1}$ from available information to precondition $\|\nabla^2 f(x_k) P_{k-1} y + \nabla f(x_k)\|$.

Preconditioned Newton's Method

Input: $P_0 = I$

for $k = 1, 2, \ldots$ **do**

`Proxy solve` $\nabla^2 f(x_k) P_{k-1} d_k = -\nabla f(x_k)$

Preconditioned Newton's Method

Input: $P_0 = I$
for $k = 1, 2, \ldots$ **do**
    `Proxy solve` $\nabla^2 f(x_k) P_{k-1} d_k = -\nabla f(x_k)$;
    `Step` $x_{k+1} = P_{k-1} d_k + x_k$

Preconditioned Newton's Method

Input: $P_0 = I$

for $k = 1, 2, \ldots$ **do**

`Proxy solve` $\nabla^2 f(x_k) P_{k-1} d_k = -\nabla f(x_k)$;

`Step` $x_{k+1} = P_{k-1} d_k + x_k$;

Calculate P_k from P_{k-1} and $\nabla^2 f(x_k) \mathcal{S}_k$.

Preconditioned Newton's Method

Input: $P_0 = I$
for $k = 1, 2, \ldots$ **do**
    `Proxy solve` $\nabla^2 f(x_k) P_{k-1} d_k = -\nabla f(x_k)$;
    `Step` $x_{k+1} = P_{k-1} d_k + x_k$;
    Calculate P_k from P_{k-1} and $\nabla^2 f(x_k)\mathcal{S}_k$.
end

Just the Facts for calculating P_k

- $\|\nabla^2 f(x_k)^{-1} - \nabla^2 f(x_{k-1})^{-1}\|$ small $\Rightarrow$ make $\|P_k - P_{k-1}\|$ small

Preconditioned Newton's Method

Input: $P_0 = I$
for $k = 1, 2, \ldots$ **do**
    `Proxy solve` $\nabla^2 f(x_k) P_{k-1} d_k = -\nabla f(x_k)$;
    `Step` $x_{k+1} = P_{k-1} d_k + x_k$;
    Calculate P_k from P_{k-1} and $\nabla^2 f(x_k) \mathcal{S}_k$.
end

Just the Facts for calculating P_k

- $\|\nabla^2 f(x_k)^{-1} - \nabla^2 f(x_{k-1})^{-1}\|$ small $\Rightarrow$ make $\|P_k - P_{k-1}\|$ small
- $\nabla^2 f(x_k)^{-1}$ is symmetric $\Rightarrow$ make P_k symmetric

Preconditioned Newton's Method

Input: $P_0 = I$
for $k = 1, 2, \ldots$ **do**
  `Proxy solve` $\nabla^2 f(x_k) P_{k-1} d_k = -\nabla f(x_k)$;
  `Step` $x_{k+1} = P_{k-1} d_k + x_k$;
  Calculate P_k from P_{k-1} and $\nabla^2 f(x_k)\mathcal{S}_k$.
end

Just the Facts for calculating P_k

- $\|\nabla^2 f(x_k)^{-1} - \nabla^2 f(x_{k-1})^{-1}\|$ small $\Rightarrow$ make $\|P_k - P_{k-1}\|$ small
- $\nabla^2 f(x_k)^{-1}$ is symmetric $\Rightarrow$ make P_k symmetric
- Has the action $\nabla^2 f(x_k)^{-1} \left(\nabla^2 f(x_k)\mathcal{S}_k\right) = \mathcal{S}_k \Rightarrow$ make $P_k \left(\nabla^2 f(x_k)\mathcal{S}_k\right) = \mathcal{S}_k$

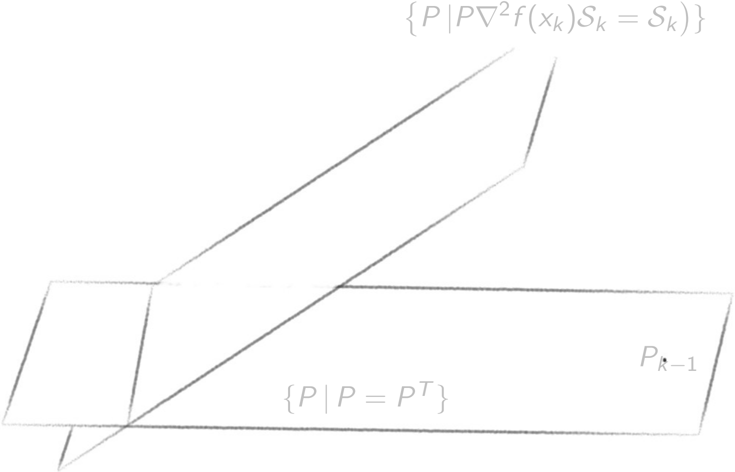
Preconditioned Newton's Method

Input: $P_0 = I$
for $k = 1, 2, \ldots$ **do**
  `Proxy solve` $\nabla^2 f(x_k) P_{k-1} d_k = -\nabla f(x_k)$;
  `Step` $x_{k+1} = P_{k-1} d_k + x_k$;
  Calculate P_k from P_{k-1} and $\nabla^2 f(x_k)\mathcal{S}_k$.
end

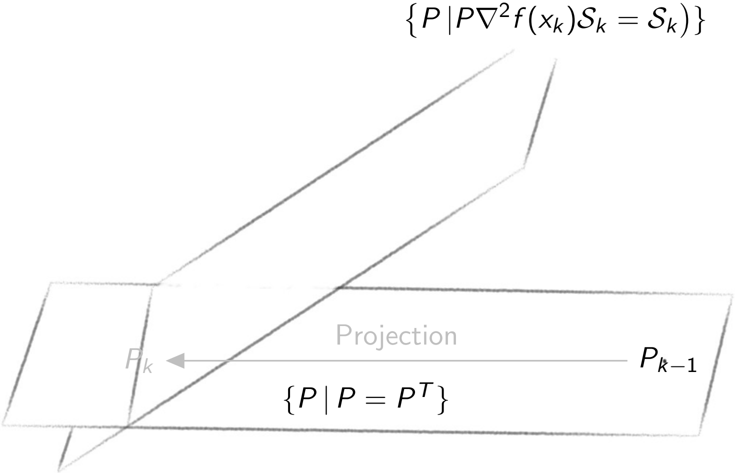
Just the Facts for calculating P_k

- $\|\nabla^2 f(x_k)^{-1} - \nabla^2 f(x_{k-1})^{-1}\|$ small $\Rightarrow$ make $\|P_k - P_{k-1}\|$ small
- $\nabla^2 f(x_k)^{-1}$ is symmetric $\Rightarrow$ make P_k symmetric
- Has the action $\nabla^2 f(x_k)^{-1} \left(\nabla^2 f(x_k)\mathcal{S}_k\right) = \mathcal{S}_k \Rightarrow$ make $P_k \left(\nabla^2 f(x_k)\mathcal{S}_k\right) = \mathcal{S}_k$

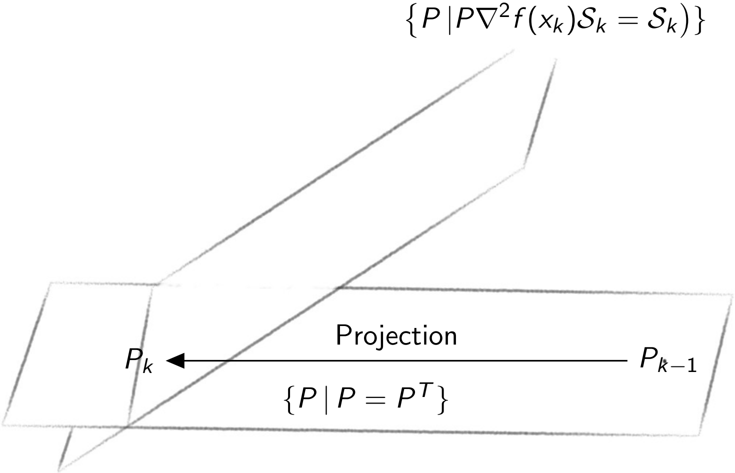
Problem: We have $P_{k-1} \approx \nabla^2 f(x_{k-1})$ we observe $\nabla^2 f(x_k)\mathcal{S}_k$ how to estimate $\nabla^2 f(x_k)^{-1}$



Problem: We have $P_{k-1} \approx \nabla^2 f(x_{k-1})$ we observe $\nabla^2 f(x_k)\mathcal{S}_k$ how to estimate $\nabla^2 f(x_k)^{-1}$



Problem: We have $P_{k-1} \approx \nabla^2 f(x_{k-1})$ we observe $\nabla^2 f(x_k)\mathcal{S}_k$ how to estimate $\nabla^2 f(x_k)^{-1}$



Hunting the Inverse

$$
\begin{aligned}
\min_{P_k} \quad & \|P_k - P_{k-1}\|^2_{Frobenius(\mathcal{W}_k)} \\
& P_k \nabla^2 f_k \mathcal{S}_k = \mathcal{S}_k \\
& P_k = P_k^T.
\end{aligned}
$$

- Iteratively updating metric; changes “slowly”
- Same action of $\nabla^2 f(x_k)^{-1}$ and P_k over $\nabla^2 f(x_k)\mathcal{S}_k$.

Hunting the Inverse

$$\min_{P_k} \quad \|P_k - P_{k-1}\|^2_{Frobenius(\mathcal{W}_k)}$$

$$P_k \nabla^2 f_k \mathcal{S}_k = \mathcal{S}_k$$

$$P_k = P_k^T.$$

- Iteratively updating metric; changes "slowly"
- Same action of $\nabla^2 f(x_k)^{-1}$ and P_k over $\nabla^2 f(x_k)\mathcal{S}_k$.
- Must be symmetric

Hunting the Inverse

$$\min_{P_k} \quad \|P_k - P_{k-1}\|^2_{Frobenius(\mathcal{W}_k)}$$
$$P_k \nabla^2 f_k \mathcal{S}_k = \mathcal{S}_k$$
$$P_k = P_k^T.$$

- Iteratively updating metric; changes "slowly"
- Same action of $\nabla^2 f(x_k)^{-1}$ and P_k over $\nabla^2 f(x_k)\mathcal{S}_k$.
- Must be symmetric

$$P_k = \nabla^2 f(x_k) + \left(I - \mathcal{W}_k \mathrm{proj}_{\mathcal{S}_k}^{\mathcal{W}_k}\right)\left(P_{k-1} - \nabla^2 f(x_k)\right)\left(I - \mathrm{proj}_{\mathcal{S}_k}^{\mathcal{W}_k} \mathcal{W}_k\right)$$

Hunting the Inverse

$$\min_{P_k} \quad \|P_k - P_{k-1}\|^2_{Frobenius(\mathcal{W}_k)}$$
$$P_k \nabla^2 f_k \mathcal{S}_k = \mathcal{S}_k$$
$$P_k = P_k^T.$$

- Iteratively updating metric; changes "slowly"
- Same action of $\nabla^2 f(x_k)^{-1}$ and P_k over $\nabla^2 f(x_k)\mathcal{S}_k$.
- Must be symmetric

$$P_k = \nabla^2 f(x_k) + \left(I - \mathcal{W}_k \mathrm{proj}_{\mathcal{S}_k}^{\mathcal{W}_k}\right)\left(P_{k-1} - \nabla^2 f(x_k)\right)\left(I - \mathrm{proj}_{\mathcal{S}_k}^{\mathcal{W}_k} \mathcal{W}_k\right)$$

Hunting the Inverse

$$\min_{P_k} \quad \boxed{\|P_k - P_{k-1}\|^2_{Frobenius(\mathcal{W}_k)}}$$
$$P_k \nabla^2 f_k \mathcal{S}_k = \mathcal{S}_k$$
$$\boxed{P_k = P_k^T.}$$

- Iteratively updating metric; changes "slowly"
- Same action of $\nabla^2 f(x_k)^{-1}$ and P_k over $\nabla^2 f(x_k)\mathcal{S}_k$.
- Must be symmetric

$$P_k = \nabla^2 f(x_k) + \left(I - \mathcal{W}_k \mathrm{proj}_{\mathcal{S}_k}^{\mathcal{W}_k}\right)\left(P_{k-1} - \nabla^2 f(x_k)\right)\left(I - \mathrm{proj}_{\mathcal{S}_k}^{\mathcal{W}_k} \mathcal{W}_k\right)$$

$\mathrm{proj}_{\mathcal{S}}^{A} A := \mathcal{S}(\mathcal{S}^T A \mathcal{S})^{-1}\mathcal{S}^T A = A-$projection onto $\mathrm{span}(\mathcal{S})$.

Hunting the Inverse

$$\min_{P_k} \quad \|P_k - P_{k-1}\|^2_{Frobenius(\mathcal{W}_k)}$$

$$P_k \nabla^2 f_k \mathcal{S}_k = \mathcal{S}_k$$

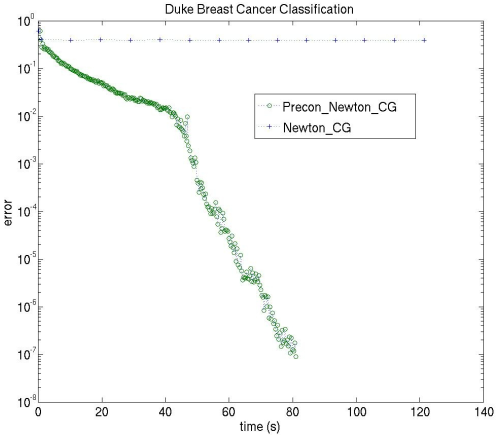
$$P_k = P_k^T.$$

- Iteratively updating metric; changes "slowly"
- Same action of $\nabla^2 f(x_k)^{-1}$ and P_k over $\nabla^2 f(x_k)\mathcal{S}_k$.
- Must be symmetric

$$P_k = \nabla^2 f(x_k) + \left(I - \mathcal{W}_k \text{proj}_{\mathcal{S}_k}^{\mathcal{W}_k}\right)(P_{k-1} - \nabla^2 f(x_k))\left(I - \text{proj}_{\mathcal{S}_k}^{\mathcal{W}_k}\mathcal{W}_k\right)$$

$\text{proj}_{\mathcal{S}}^{A} A := \mathcal{S}(\mathcal{S}^T A \mathcal{S})^{-1}\mathcal{S}^T A = A-$projection onto $\text{span}(\mathcal{S})$.

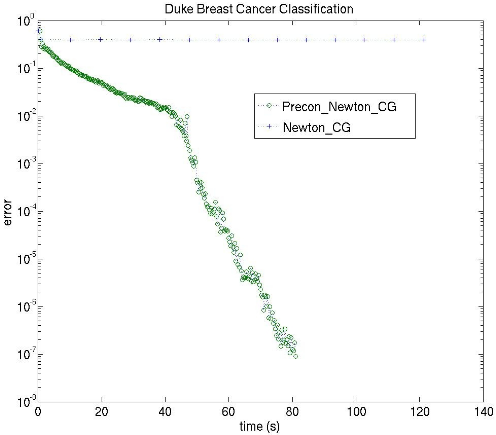
Testing on Duke-Breast-Cancer Classification



7129 features and 44 data

Preconditioning can make all the difference

Testing on Duke-Breast-Cancer Classification



7129 features and 44 data

Preconditioning can make all the difference

References

Gower, Robert M., Gondzio, Jacek (2014).
Action constrained quasi-Newton methods (in progress)

Fletcher, B. R., Powell, M. J. D. (1960).
A rapidly convergent descent method for minimization.

Davidon, W. C. (1959). Variable metric method for minimization.

Goldfarb, D. (1970).
A Family of Variable-Metric Methods Derived by Variational Means.
Mathematics of Computation, 24(109), 23.