Action constrained quasi-Newton methods

Robert Gower

EUROPT 2014 Workshop on Advances in Continuous Optimization

July 10, 2014

What and why quasi-Newton?

Classic quasi-Newton

New Action Constrained quasi-Newton

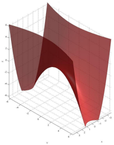
Implementing a Preconditioned Newton-CG with new metric

What's new

- A few Hessian-vector products are cheap $\Rightarrow$ use a handful to build Hessian approximation [1]
- Framework for "tracking" inverses of matrix fields
- General purpose Newton-CG preconditioners
- Good results on regularized logistic regression (compared to BFGS or Newton-CG)

[1] Walther, A. (2008). Computing sparse Hessians with automatic differentiation. ACM Trans. Math. Software, 34(1), Art. 3, 15.

Solving sequences of linear systems

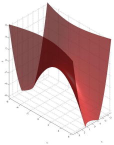


$$\min \quad f(x) \quad (C^2 - \text{diffeomorphism})$$

$$f(x_k + d) \approx f_k + \nabla f_k^T d + \tfrac{1}{2} d^T \nabla^2 f_k d$$

x_0

Solving sequences of linear systems

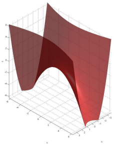


$$\min \quad f(x)$$

$$f(x_k + d) \quad \approx f_k + \nabla f_k^T d + \frac{1}{2} d^T \nabla^2 f_k d$$

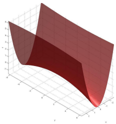
x_0

Solving sequences of linear systems



$$\min \quad f(x)$$

$$f(x_k + d) \approx f_k + \nabla f_k^T d + \frac{1}{2} d^T \nabla^2 f_k d$$

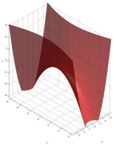


x_0

$d_0 = -\nabla^2 f_0^{-1} \nabla f_0$

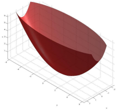
x_1

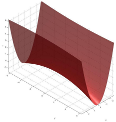
Solving sequences of linear systems



$$\min \quad f(x)$$

$$f(x_k + d) \quad \approx f_k + \nabla f_k^T d + \tfrac{1}{2} d^T \nabla^2 f_k d$$





x_0

$d_0 = -\nabla^2 f_0^{-1} \nabla f_0$

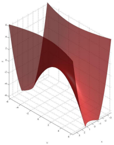
x_1

x_2

$d_1 = -\nabla^2 f_1^{-1} \nabla f_1$

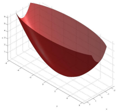
The quadratic model "slowly" changes; how to take advantage?

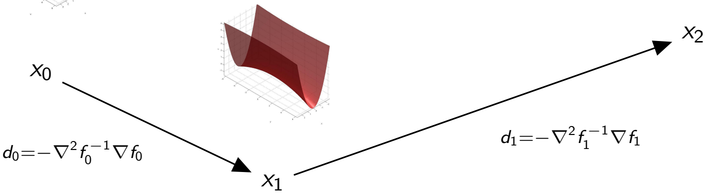
Solving sequences of linear systems



$$\min \quad f(x)$$

$$f(x_k + d) \quad \approx f_k + \nabla f_k^T d + \tfrac{1}{2} d^T \nabla^2 f_k d$$





The quadratic model "slowly" changes; how to take advantage?

Maintain an approximation of the inverse: $H_{k+1} \approx \nabla^2 f_{k+1}^{-1}$.

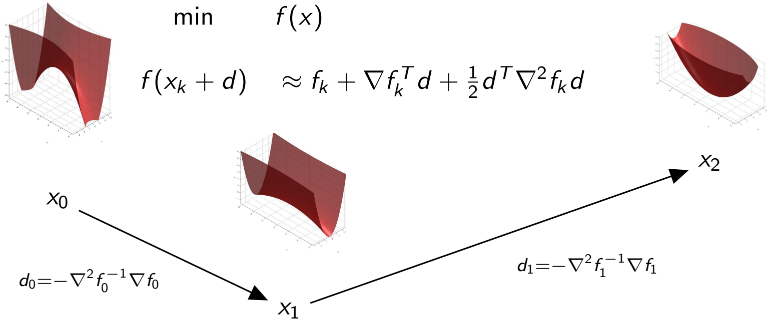
"Slowly" update with cheap low rank matrices: $H_{k+1} = H_k + E_k$.

H_k is a *metric* matrix.

Solving sequences of linear systems

$$\min \quad f(x)$$

$$f(x_k + d) \quad \approx f_k + \nabla f_k^T d + \tfrac{1}{2} d^T \nabla^2 f_k d$$



The quadratic model "slowly" changes; how to take advantage?
Maintain an approximation of the inverse: $H_{k+1} \approx \nabla^2 f_{k+1}^{-1}$.
"Slowly" update with cheap low rank matrices: $H_{k+1} = H_k + E_k$.
H_k is a *metric* matrix.

Classic approach: imitating action of Hessian with the secant equation

$$\mathcal{A}\nabla^2 f_{k+1} = \int_0^1 \nabla^2 f(x_k + td_k)dt, \quad \gamma_k = (\nabla f_{k+1} - \nabla f_k).$$

Fundamental Theorem of Calculus says

$$\mathcal{A}\nabla^2 f_{k+1} \cdot d_k = \gamma_k \Rightarrow$$
$$d_k = (\mathcal{A}\nabla^2 f_{k+1})^{-1}\gamma_k \approx \left(\nabla^2 f_{k+1}\right)^{-1}\gamma_k$$

[2] Fletcher, B. R., & Powell, M. J. D. (1960). A rapidly convergent descent method for minimization, (1).

Classic approach: imitating action of Hessian with the secant equation

$$\mathcal{A}\nabla^2 f_{k+1} = \int_0^1 \nabla^2 f(x_k + td_k)dt, \quad \gamma_k = (\nabla f_{k+1} - \nabla f_k).$$

Fundamental Theorem of Calculus says

$$\mathcal{A}\nabla^2 f_{k+1} \cdot d_k = \gamma_k \Rightarrow$$
$$d_k = (\mathcal{A}\nabla^2 f_{k+1})^{-1}\gamma_k \approx \left(\nabla^2 f_{k+1}\right)^{-1}\gamma_k$$

Choose a metric that satisfies

The Secant Equation

$$d_k = H_{k+1}\gamma_k = H_{k+1}\mathcal{A}\nabla^2 f_{k+1} d_k$$

Still under-determined. Least squares idea + slowly changing [2]

Classic approach: imitating action of Hessian with the secant equation

$$\mathcal{A}\nabla^2 f_{k+1} = \int_0^1 \nabla^2 f(x_k + td_k)dt, \quad \gamma_k = (\nabla f_{k+1} - \nabla f_k).$$

Fundamental Theorem of Calculus says

$$\begin{aligned} \mathcal{A}\nabla^2 f_{k+1} \cdot d_k = \gamma_k \Rightarrow& \\ & d_k = (\mathcal{A}\nabla^2 f_{k+1})^{-1}\gamma_k \approx \left(\nabla^2 f_{k+1}\right)^{-1}\gamma_k \end{aligned}$$

Choose a metric that satisfies

The Secant Equation

$$d_k = H_{k+1}\gamma_k = H_{k+1}\mathcal{A}\nabla^2 f_{k+1} d_k$$

Still under-determined. Least squares idea + slowly changing [2]

[2] Fletcher, B. R., & Powell, M. J. D. (1960). A rapidly convergent descent method for minimization, (1).

Classic approach: imitating action of Hessian with the secant equation

$$\mathcal{A}\nabla^2 f_{k+1} = \int_0^1 \nabla^2 f(x_k + td_k)dt, \quad \gamma_k = (\nabla f_{k+1} - \nabla f_k).$$

Fundamental Theorem of Calculus says

$$\begin{aligned} \mathcal{A}\nabla^2 f_{k+1} \cdot d_k = \gamma_k \Rightarrow& \\ & d_k = (\mathcal{A}\nabla^2 f_{k+1})^{-1}\gamma_k \approx \left(\nabla^2 f_{k+1}\right)^{-1}\gamma_k \end{aligned}$$

Choose a metric that satisfies

The Secant Equation

$$d_k = H_{k+1}\gamma_k = H_{k+1}\mathcal{A}\nabla^2 f_{k+1} d_k$$

Still under-determined. Least squares idea + slowly changing [2]

[2] Fletcher, B. R., & Powell, M. J. D. (1960). A rapidly convergent descent method for minimization, (1).

Designing classic quasi-Newton

$$\min_{H_{k+1}} \quad \|H_{k+1} - H_k\|^2_{Frobenius(W)} = \|E_k\|^2_{Frobenius(W)} = \sum_{i,j} E^2_{i,j} W^2_{i,j}$$

$$\text{s.t.} \quad H_{k+1}\mathcal{A}\nabla^2 f_{k+1} d_k = d_k, \qquad H_{k+1} = H^T_{k+1}.$$

- Iteratively updating metric; changes "slowly"[3]

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

Designing classic quasi-Newton

$$\min_{H_{k+1}} \quad \|H_{k+1} - H_k\|^2_{Frobenius(W)} = \|E_k\|^2_{Frobenius(W)} = \sum_{i,j} E^2_{i,j} W^2_{i,j}$$

$$\text{s.t.} \quad H_{k+1}\mathcal{A}\nabla^2 f_{k+1} d_k = d_k, \qquad H_{k+1} = H^T_{k+1}.$$

- Iteratively updating metric; changes "slowly"[3]
- Secant equation Must be symmetric

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

Designing classic quasi-Newton

$$\min_{H_{k+1}} \quad \|H_{k+1} - H_k\|^2_{Frobenius(W)} = \|E_k\|^2_{Frobenius(W)} = \sum_{i,j} E_{i,j}^2 W_{i,j}^2$$

$$\text{s.t.} \quad H_{k+1}\mathcal{A}\nabla^2 f_{k+1} d_k = d_k, \qquad H_{k+1} = H_{k+1}^T.$$

- Iteratively updating metric; changes "slowly"[3]
- Secant equation Must be symmetric

BFGS by choosing $W = \mathcal{A}\nabla^2 f_{k+1}$,

$$H_{k+1} = \frac{d_k d_k^T}{d_k^T \gamma_k} + \left(I - \frac{d_k \gamma_k^T}{d_k^T \gamma_k}\right) H_k \left(I - \frac{\gamma_k d_k^T}{d_k^T \gamma_k}\right).$$

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

Designing classic quasi-Newton

$$\min_{H_{k+1}} \quad \|H_{k+1} - H_k\|^2_{Frobenius(W)} = \|E_k\|^2_{Frobenius(W)} = \textstyle\sum_{i,j} E^2_{i,j} W^2_{i,j}$$

$$\text{s.t.} \quad H_{k+1}\mathcal{A}\nabla^2 f_{k+1} d_k = d_k, \qquad H_{k+1} = H^T_{k+1}.$$

- Iteratively updating metric; changes "slowly"[3]
- Secant equation Must be symmetric

BFGS by choosing $W = \mathcal{A}\nabla^2 f_{k+1}$,

$$H_{k+1} = \frac{d_k d_k^T}{d_k^T \gamma_k} + \left(I - \frac{d_k \gamma_k^T}{d_k^T \gamma_k}\right) H_k \left(I - \frac{\gamma_k d_k^T}{d_k^T \gamma_k}\right).$$

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

Designing classic quasi-Newton

$$\min_{H_{k+1}} \quad \|H_{k+1} - H_k\|^2_{Frobenius(W)} = \|E_k\|^2_{Frobenius(W)} = \sum_{i,j} E^2_{i,j} W^2_{i,j}$$

$$\text{s.t.} \quad H_{k+1}\mathcal{A}\nabla^2 f_{k+1} d_k = d_k, \qquad H_{k+1} = H^T_{k+1}.$$

- Iteratively updating metric; changes "slowly"[3]
- Secant equation Must be symmetric

BFGS by choosing $W = \mathcal{A}\nabla^2 f_{k+1}$,

$$H_{k+1} = \text{proj}_{d_k}^{\mathcal{A}\nabla^2 f_k} + \left(I - \text{proj}_{d_k}^{\mathcal{A}\nabla^2 f_k}\mathcal{A}\nabla^2 f_k\right) H_k \left(I - \mathcal{A}\nabla^2 f_k \text{proj}_{d_k}^{\mathcal{A}\nabla^2 f_k}\right).$$

$\text{proj}_d^A A := d(d^T A d)^{-1} d^T A =$ oblique $A-$projection onto span(d).

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

Designing classic quasi-Newton

$$\min_{H_{k+1}} \quad \|H_{k+1} - H_k\|^2_{Frobenius(W)} = \|E_k\|^2_{Frobenius(W)} = \sum_{i,j} E^2_{i,j} W^2_{i,j}$$

$$\text{s.t.} \quad H_{k+1}\mathcal{A}\nabla^2 f_{k+1} d_k = d_k, \qquad H_{k+1} = H^T_{k+1}.$$

- Iteratively updating metric; changes “slowly”[3]
- Secant equation   Must be symmetric

BFGS by choosing $W = \mathcal{A}\nabla^2 f_{k+1}$,

$$H_{k+1} = \text{proj}_{d_k}^{\mathcal{A}\nabla^2 f_k} + \left(I - \text{proj}_{d_k}^{\mathcal{A}\nabla^2 f_k}\mathcal{A}\nabla^2 f_k\right) H_k \left(I - \mathcal{A}\nabla^2 f_k \text{proj}_{d_k}^{\mathcal{A}\nabla^2 f_k}\right).$$

$\text{proj}_d^A A := d(d^T A d)^{-1} d^T A =$oblique $A-$projection onto $\text{span}(d)$.

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

Designing classic quasi-Newton

$$\min_{H_{k+1}} \quad \|H_{k+1} - H_k\|^2_{Frobenius(W)} = \|E_k\|^2_{Frobenius(W)} = \sum_{i,j} E^2_{i,j} W^2_{i,j}$$

$$\text{s.t.} \quad H_{k+1}\mathcal{A}\nabla^2 f_{k+1} d_k = d_k, \qquad H_{k+1} = H^T_{k+1}.$$

- Iteratively updating metric; changes "slowly"[3]
- Secant equation Must be symmetric

BFGS by choosing $W = \mathcal{A}\nabla^2 f_{k+1}$,

$$H_{k+1} = \mathrm{proj}_{d_k}^{\mathcal{A}\nabla^2 f_k} + \left(I - \mathrm{proj}_{d_k}^{\mathcal{A}\nabla^2 f_k}\mathcal{A}\nabla^2 f_k\right) H_k \left(I - \mathcal{A}\nabla^2 f_k \mathrm{proj}_{d_k}^{\mathcal{A}\nabla^2 f_k}\right).$$

$\mathrm{proj}_d^A A := d(d^T A d)^{-1} d^T A =$oblique $A-$projection onto $\mathrm{span}(d)$.

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

Designing the action constrained metrics

$$\min_{H_{k+1}} \quad \|H_{k+1} - H_k\|^2_{Frobenius(W)}$$

$$\text{s.t.} \quad H_{k+1}\nabla^2 f_{k+1} D_k = D_k, \qquad H_{k+1} = H^T_{k+1}.$$

- Iteratively updating metric; changes "slowly"
- Same action of $\nabla^2 f^{-1}_{k+1}$ and H_{k+1} over $\nabla^2 f_{k+1} D_k$ where $D_k \in \mathbb{R}^{n\times p}, p << n$ is a tall thin matrix . Must be symmetric

$$H_{k+1} = H_k + W^{-1}\text{proj}^{W^{-1}}_{\nabla^2 f_{k+1} D_k}(I - H_k\nabla^2 f_{k+1})\left(I - \text{proj}^{W^{-1}}_{\nabla^2 f_{k+1} D_k} W^{-1}\right)$$
$$+ (I - H_k\nabla^2 f_{k+1})\text{proj}^{W^{-1}}_{\nabla^2 f_{k+1} D_k} W^{-1}$$

Designing the action constrained metrics

$$\min_{H_{k+1}} \quad \|H_{k+1} - H_k\|^2_{Frobenius(W)}$$

$$\text{s.t.} \quad H_{k+1}\nabla^2 f_{k+1} D_k = D_k, \qquad H_{k+1} = H_{k+1}^T.$$

- Iteratively updating metric; changes "slowly"
- Same action of $\nabla^2 f_{k+1}^{-1}$ and H_{k+1} over $\nabla^2 f_{k+1} D_k$ where $D_k \in \mathbb{R}^{n\times p}, p << n$ is a tall thin matrix . Must be symmetric

$$H_{k+1} = H_k + W^{-1}\text{proj}^{W^{-1}}_{\nabla^2 f_{k+1} D_k}(I - H_k\nabla^2 f_{k+1})\left(I - \text{proj}^{W^{-1}}_{\nabla^2 f_{k+1} D_k} W^{-1}\right)$$
$$+ (I - H_k\nabla^2 f_{k+1})\text{proj}^{W^{-1}}_{\nabla^2 f_{k+1} D_k} W^{-1}$$

$\text{proj}^A_D A := D(D^T A D)^{-1} D^T A.$

Designing the action constrained metrics

$$\min_{H_{k+1}} \quad \|H_{k+1} - H_k\|^2_{Frobenius(W)}$$

$$\text{s.t.} \quad H_{k+1}\nabla^2 f_{k+1} D_k = D_k, \quad H_{k+1} = H_{k+1}^T.$$

- Iteratively updating metric; changes "slowly"
- Same action of $\nabla^2 f_{k+1}^{-1}$ and H_{k+1} over $\nabla^2 f_{k+1} D_k$ where $D_k \in \mathbb{R}^{n\times p}, p << n$ is a tall thin matrix . Must be symmetric

$$H_{k+1} = H_k + W^{-1}\mathrm{proj}^{W^{-1}}_{\nabla^2 f_{k+1} D_k}(I - H_k\nabla^2 f_{k+1})\left(I - \mathrm{proj}^{W^{-1}}_{\nabla^2 f_{k+1} D_k} W^{-1}\right)$$
$$+ (I - H_k\nabla^2 f_{k+1})\mathrm{proj}^{W^{-1}}_{\nabla^2 f_{k+1} D_k} W^{-1}$$

$$\mathrm{proj}^A_D A := D(D^T A D)^{-1} D^T A.$$

Reverse Automatic Differentiation $\nabla^2 f_{k+1}(D_k)$ costs $O(p)$!

Designing the action constrained metrics

$$\min_{H_{k+1}} \quad \|H_{k+1} - H_k\|^2_{Frobenius(W)}$$

$$\text{s.t.} \quad H_{k+1}\nabla^2 f_{k+1} D_k = D_k, \qquad H_{k+1} = H_{k+1}^T.$$

- Iteratively updating metric; changes "slowly"
- Same action of $\nabla^2 f_{k+1}^{-1}$ and H_{k+1} over $\nabla^2 f_{k+1} D_k$ where $D_k \in \mathbb{R}^{n\times p}, p << n$ is a tall thin matrix . Must be symmetric

$$H_{k+1} = H_k + W^{-1}\text{proj}^{W^{-1}}_{\nabla^2 f_{k+1} D_k}(I - H_k\nabla^2 f_{k+1})\left(I - \text{proj}^{W^{-1}}_{\nabla^2 f_{k+1} D_k} W^{-1}\right)$$
$$+ (I - H_k\nabla^2 f_{k+1})\text{proj}^{W^{-1}}_{\nabla^2 f_{k+1} D_k} W^{-1}$$

$\text{proj}^A_D A := D(D^TAD)^{-1}D^TA.$

Reverse Automatic Differentiation $\nabla^2 f_{k+1}(D_k)$ costs $O(p)$!

A small rank$-3p$ update.

Designing the action constrained metrics

$$\min_{H_{k+1}} \quad \|H_{k+1} - H_k\|^2_{Frobenius(W)}$$

$$\text{s.t.} \quad H_{k+1}\nabla^2 f_{k+1} D_k = D_k, \qquad H_{k+1} = H_{k+1}^T.$$

- Iteratively updating metric; changes "slowly"
- Same action of $\nabla^2 f_{k+1}^{-1}$ and H_{k+1} over $\nabla^2 f_{k+1} D_k$ where $D_k \in \mathbb{R}^{n\times p}, p << n$ is a tall thin matrix . Must be symmetric

$$H_{k+1} = H_k + W^{-1}\text{proj}^{W^{-1}}_{\nabla^2 f_{k+1}D_k}(I - H_k\nabla^2 f_{k+1})\left(I - \text{proj}^{W^{-1}}_{\nabla^2 f_{k+1}D_k} W^{-1}\right)$$
$$+ (I - H_k\nabla^2 f_{k+1})\text{proj}^{W^{-1}}_{\nabla^2 f_{k+1}D_k} W^{-1}$$

$\text{proj}^A_D A := D(D^T AD)^{-1}D^T A.$

Reverse Automatic Differentiation $\nabla^2 f_{k+1}(D_k)$ costs $O(p)$!

A small rank$-3p$ update. $W =?$ and $D_k =?$

Designing the action constrained metrics

$$\min_{H_{k+1}} \quad \|H_{k+1} - H_k\|^2_{Frobenius(W)}$$
$$\text{s.t.} \quad H_{k+1}\nabla^2 f_{k+1} D_k = D_k, \quad H_{k+1} = H_{k+1}^T.$$

- Iteratively updating metric; changes "slowly"
- Same action of $\nabla^2 f_{k+1}^{-1}$ and H_{k+1} over $\nabla^2 f_{k+1} D_k$ where $D_k \in \mathbb{R}^{n\times p}, p << n$ is a tall thin matrix . Must be symmetric

$$H_{k+1} = H_k + W^{-1}\mathrm{proj}^{W^{-1}}_{\nabla^2 f_{k+1} D_k}(I - H_k\nabla^2 f_{k+1})\left(I - \mathrm{proj}^{W^{-1}}_{\nabla^2 f_{k+1} D_k} W^{-1}\right)$$
$$+ (I - H_k\nabla^2 f_{k+1})\mathrm{proj}^{W^{-1}}_{\nabla^2 f_{k+1} D_k} W^{-1}$$

$\mathrm{proj}^A_D A := D(D^T A D)^{-1} D^T A.$

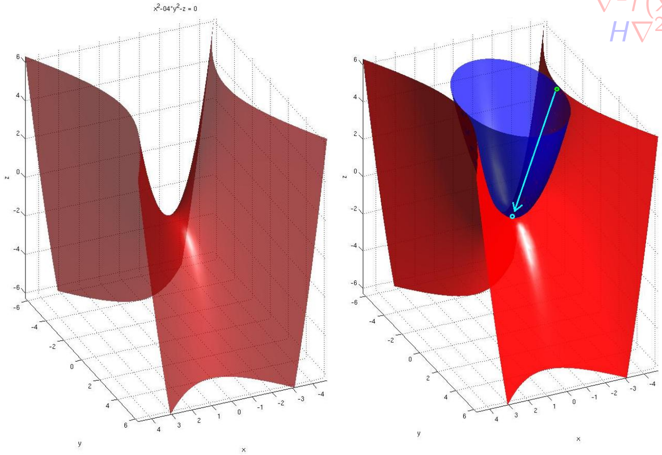
Reverse Automatic Differentiation $\nabla^2 f_{k+1}(D_k)$ costs $O(p)$!

A small rank$-3p$ update. $W = ?$ and $D_k = ?$

Overall idea of D_k

Build metric H that captures $D = [d_1, \ldots, d_p]$ directions of positive curvature

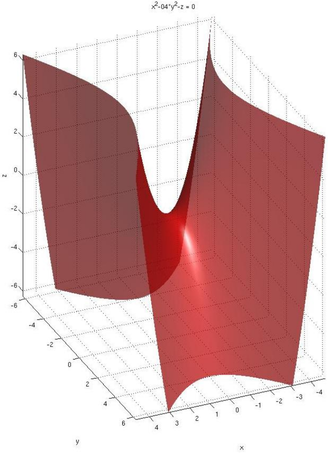
$$\nabla^2 f(x) D = H^{-1} D$$
$$H \nabla^2 f(x) D = D$$

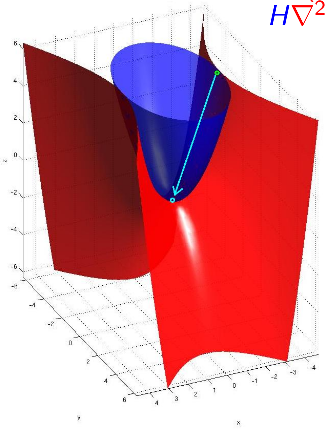


Overall idea of D_k

Build metric H that captures $D = [d_1, \ldots, d_p]$ directions of positive curvature

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

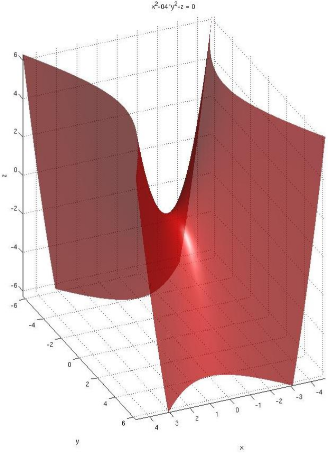


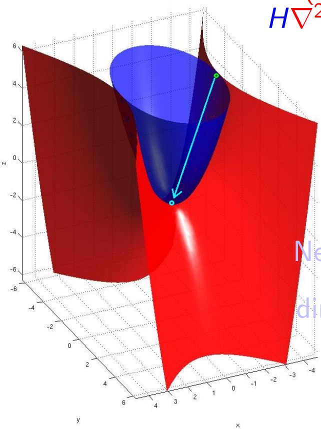


Overall idea of D_k

Build metric H that captures $D = [d_1, \ldots, d_p]$ directions of positive curvature

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

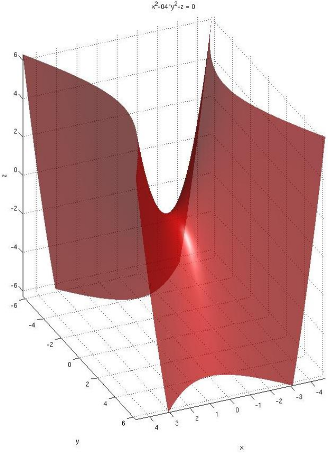


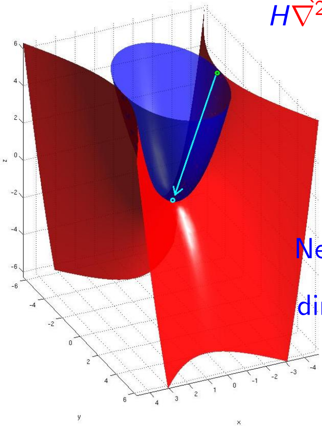


Newton-CG produces directions of positive curvature!

Overall idea of D_k

Build metric H that captures $D = [d_1, \ldots, d_p]$ directions of positive curvature





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

Newton-CG produces directions of positive curvature!

Choosing D_k and W

Analogous to BFGS by choosing $W = \nabla^2 f_k$

quNac: quasi-Newton action constrained

$$H_{k+1} = \mathrm{proj}_{D_k}^{\nabla^2 f_{k+1}} + (I - \mathrm{proj}_{D_k}^{\nabla^2 f_{k+1}} \nabla^2 f_{k+1}) H_k (I - \nabla^2 f_{k+1} \mathrm{proj}_{D_k}^{\nabla^2 f_{k+1}}).$$

- A rank$-2p$ update
- Only need to calculate the p columns $\nabla^2 f_{k+1}(D_k)$.
- $\mathrm{proj}_{D_k}^{\nabla^2 f_{k+1}} = D_k(D_k^T \nabla^2 f_{k+1} D_k)^{-1} D_k^T$ is expensive?

Choose D_k according to

- **Hereditary** property $\Rightarrow$ for local convergence
- **Descent property** $\Rightarrow$ for Global stability

Quadratic Hereditary $\nabla^2 f(x) \equiv Q$ property

If $[D_1, \ldots, D_k] \in \mathbb{R}^{n \times n}$ and $D_i^T Q D_j = 0$ for $1 \leq j < i \leq k$ then

$$H_{k+1} Q D_i = D_i, \quad \text{for } i = 1, \ldots, k.$$

Lemma: $H_{k+1} Q = I \Rightarrow H_{k+1} = Q^{-1}$.

Quadratic Hereditary $\nabla^2 f(x) \equiv Q$ property

If $[D_1, \ldots, D_k] \in \mathbb{R}^{n\times n}$ and $D_i^T Q D_j = 0$ for $1 \leq j < i \leq k$ then

$$H_{k+1} Q D_i = D_i, \quad \text{for } i = 1, \ldots, k.$$

Lemma: $H_{k+1} Q = I \Rightarrow H_{k+1} = Q^{-1}$.

proof:

$$\begin{aligned} H_{k+1} &= H_k + E_k \\ &= H_i + E_i + \cdots + E_k. \end{aligned}$$

If $D_i^T Q D_j = 0$ for $j < i$ then

$$\begin{aligned} &\left(\text{proj}_{D_j}^{Q} Q D_i = 0\right) \Rightarrow \left(E_j Q D_i = 0\right). \\ &\Rightarrow H_{k+1} Q D_i = H_{i+1} Q D_i = D_i. \quad \square \end{aligned}$$

Choose D_k conjugate to $\nabla^2 f_{k+1}$ as an approximation!

Quadratic Hereditary $\nabla^2 f(x) \equiv Q$ property

If $[D_1, \ldots, D_k] \in \mathbb{R}^{n\times n}$ and $D_i^T Q D_j = 0$ for $1 \le j < i \le k$ then

$$H_{k+1} Q D_i = D_i, \quad \text{for } i = 1, \ldots, k.$$

Lemma: $H_{k+1} Q = I \Rightarrow H_{k+1} = Q^{-1}$.

proof:

$$\begin{aligned} H_{k+1} &= H_k + E_k \\ &= H_i + E_i + \cdots + E_k. \end{aligned}$$

Quadratic Hereditary $\nabla^2 f(x) \equiv Q$ property

If $[D_1, \ldots, D_k] \in \mathbb{R}^{n\times n}$ and $D_i^T Q D_j = 0$ for $1 \leq j < i \leq k$ then

$$H_{k+1} Q D_i = D_i, \quad \text{for } i = 1, \ldots, k.$$

Lemma: $H_{k+1} Q = I \Rightarrow H_{k+1} = Q^{-1}$.
proof:

$$\begin{aligned} H_{k+1} &= H_k + E_k \\ &= H_i + E_i + \cdots + E_k. \end{aligned}$$

If $D_i^T Q D_j = 0$ for $j < i$ then

Quadratic Hereditary $\nabla^2 f(x) \equiv Q$ property

If $[D_1, \ldots, D_k] \in \mathbb{R}^{n \times n}$ and $D_i^T Q D_j = 0$ for $1 \leq j < i \leq k$ then

$$H_{k+1} Q D_i = D_i, \quad \text{for } i = 1, \ldots, k.$$

Lemma: $H_{k+1} Q = I \Rightarrow H_{k+1} = Q^{-1}$.
proof:

$$\begin{aligned} H_{k+1} &= H_k + E_k \\ &= H_i + E_i + \cdots + E_k. \end{aligned}$$

If $D_i^T Q D_j = 0$ for $j < i$ then

$$\left(\text{proj}_{D_j}^{Q} Q D_i = 0\right) \Rightarrow \left(E_j Q D_i = 0\right).$$
$$\Rightarrow H_{k+1} Q D_i = H_{i+1} Q D_i = D_i. \quad \square$$

Quadratic Hereditary $\nabla^2 f(x) \equiv Q$ property

If $[D_1, \ldots, D_k] \in \mathbb{R}^{n\times n}$ and $D_i^T Q D_j = 0$ for $1 \leq j < i \leq k$ then

$$H_{k+1} Q D_i = D_i, \quad \text{for } i = 1, \ldots, k.$$

Lemma: $H_{k+1} Q = I \Rightarrow H_{k+1} = Q^{-1}$.

proof:

$$\begin{aligned} H_{k+1} &= H_k + E_k \\ &= H_i + E_i + \cdots + E_k. \end{aligned}$$

If $D_i^T Q D_j = 0$ for $j < i$ then

$$\left(\text{proj}_{D_j}^{Q} Q D_i = 0\right) \Rightarrow \left(E_j Q D_i = 0\right).$$
$$\Rightarrow H_{k+1} Q D_i = H_{i+1} Q D_i = D_i. \quad \square$$

Choose D_k conjugate to $\nabla^2 f_{k+1}$ as an approximation!

Quadratic Hereditary $\nabla^2 f(x) \equiv Q$ property

If $[D_1, \ldots, D_k] \in \mathbb{R}^{n \times n}$ and $D_i^T Q D_j = 0$ for $1 \leq j < i \leq k$ then

$$H_{k+1} Q D_i = D_i, \quad \text{for } i = 1, \ldots, k.$$

Lemma: $H_{k+1} Q = I \Rightarrow H_{k+1} = Q^{-1}$.

proof:

$$\begin{aligned} H_{k+1} &= H_k + E_k \\ &= H_i + E_i + \cdots + E_k. \end{aligned}$$

If $D_i^T Q D_j = 0$ for $j < i$ then

$$\begin{aligned} \left(\text{proj}_{D_j}^{Q} Q D_i = 0\right) &\Rightarrow \left(E_j Q D_i = 0\right). \\ &\Rightarrow H_{k+1} Q D_i = H_{i+1} Q D_i = D_i. \quad \square \end{aligned}$$

Choose D_k conjugate to $\nabla^2 f_{k+1}$ as an approximation!

Quadratic Hereditary $\nabla^2 f(x) \equiv Q$ property

If $[D_1, \ldots, D_k] \in \mathbb{R}^{n\times n}$ and $D_i^T QW^{-1}QD_j = 0$ for $1 \leq j < i \leq k$ then

$$H_{k+1}QD_i = D_i, \quad \text{for } i = 1, \ldots, k.$$

Lemma: $H_{k+1}Q = I \Rightarrow H_{k+1} = Q^{-1}$.

proof:

$$\begin{aligned} H_{k+1} &= H_k + E_k \\ &= H_i + E_i + \cdots + E_k. \end{aligned}$$

If $D_i^T QW^{-1}QD_j = 0$ for $j < i$ then

$$\begin{aligned} \left(\text{proj}^W_{QD_j} W^{-1}QD_i = 0\right) &\Rightarrow (E_jQD_i = 0) \\ &\Rightarrow H_{k+1}QD_i = H_{i+1}QD_i = D_i. \quad \square \end{aligned}$$

Choose D_k conjugate to $\nabla^2 f_{k+1}$ as an approximation!

Descent and Positive Definiteness

Descent: $d_k = -H_k \nabla f_k$ is less then 90^o with $-\nabla f_k$.

Sufficient Descent condition

If $H_k \succ 0$ then $-d_k^T \nabla f_k = \nabla f_k^T H_k \nabla f_k > 0$.

Classic quasi-Newton

If $H_k \succ 0$ and $\gamma_k^T d_k = d_k^T \mathcal{A} \nabla^2 f_{k+1} d_k > 0$ then $H_{k+1} \succ 0$.

Descent and Positive Definiteness

Descent: $d_k = -H_k \nabla f_k$ is less then 90^o with $-\nabla f_k$.

Sufficient Descent condition

If $H_k \succ 0$ then $-d_k^T \nabla f_k = \nabla f_k^T H_k \nabla f_k > 0$.

Classic quasi-Newton

If $H_k \succ 0$ and $\gamma_k^T d_k = d_k^T \mathcal{A} \nabla^2 f_{k+1} d_k > 0$ then $H_{k+1} \succ 0$.

quNac Descent condition

If $H_k \succ 0$ and D_k columns are directions of positive curvature $D_k^T \nabla^2 f_{k+1} D_k \succ 0$ then $H_{k+1} \succ 0$.

Descent and Positive Definiteness

Descent: $d_k = -H_k \nabla f_k$ is less then 90^o with $-\nabla f_k$.

Sufficient Descent condition

If $H_k \succ 0$ then $-d_k^T \nabla f_k = \nabla f_k^T H_k \nabla f_k > 0$.

Classic quasi-Newton

If $H_k \succ 0$ and $\gamma_k^T d_k = d_k^T \mathcal{A} \nabla^2 f_{k+1} d_k > 0$ then $H_{k+1} \succ 0$.

quNac Descent condition

If $H_k \succ 0$ and D_k columns are directions of positive curvature $D_k^T \nabla^2 f_{k+1} D_k \succ 0$ then $H_{k+1} \succ 0$.

Implementing a Preconditioned Newton-CG with quNac

- Call Conjugate gradients on $\nabla^2 f_{k+1} d = -\nabla f_{k+1}$ to get D_k directions of positive curvature
- $\text{proj}_{D_k}^{\nabla^2 f_{k+1}} = D_k (D_k^T \nabla^2 f_{k+1} D_k)^{-1} D_k^T$ is $O\left(p^2\right)$ because $D_k^T \nabla^2 f_{k+1} D_k =$diagonal !

Implementing a Preconditioned Newton-CG with quNac

- Call Conjugate gradients on $\nabla^2 f_{k+1} d = -\nabla f_{k+1}$ to get D_k directions of positive curvature
- $\mathrm{proj}_{D_k}^{\nabla^2 f_{k+1}} = D_k (D_k^T \nabla^2 f_{k+1} D_k)^{-1} D_k^T$ is $O\left(p^2\right)$ because $D_k^T \nabla^2 f_{k+1} D_k =$diagonal !
- H_{k+1} is positive definite and $H_{k+1} \nabla^2 f_{k+1} D_k = D_k$, thus $H_{k+1} \nabla^2 f_{k+1}$ has concentrated Eigen-values $\Rightarrow$ Good preconditioner CG!

Implementing a Preconditioned Newton-CG with quNac

- Call Conjugate gradients on $\nabla^2 f_{k+1} d = -\nabla f_{k+1}$ to get D_k directions of positive curvature
- $\text{proj}_{D_k}^{\nabla^2 f_{k+1}} = D_k (D_k^T \nabla^2 f_{k+1} D_k)^{-1} D_k^T$ is $O\left(p^2\right)$ because $D_k^T \nabla^2 f_{k+1} D_k =$diagonal !
- H_{k+1} is positive definite and $H_{k+1} \nabla^2 f_{k+1} D_k = D_k$, thus $H_{k+1} \nabla^2 f_{k+1}$ has concentrated Eigen-values $\Rightarrow$ Good preconditioner CG!
- Use $H_k \approx \nabla^2 f_{k+1}$ as a preconditioner in a Newton-CG framework!

Implementing a Preconditioned Newton-CG with quNac

- Call Conjugate gradients on $\nabla^2 f_{k+1} d = -\nabla f_{k+1}$ to get D_k directions of positive curvature
- $\text{proj}_{D_k}^{\nabla^2 f_{k+1}} = D_k (D_k^T \nabla^2 f_{k+1} D_k)^{-1} D_k^T$ is $O\left(p^2\right)$ because $D_k^T \nabla^2 f_{k+1} D_k =$diagonal !
- H_{k+1} is positive definite and $H_{k+1} \nabla^2 f_{k+1} D_k = D_k$, thus $H_{k+1} \nabla^2 f_{k+1}$ has concentrated Eigen-values $\Rightarrow$ Good preconditioner CG!
- Use $H_k \approx \nabla^2 f_{k+1}$ as a preconditioner in a Newton-CG framework!

quNac quasi-Newton action Constrained

- $H_0, x_0 \in \mathbb{R}, \; k = 0$

quNac quasi-Newton action Constrained

- $H_0, x_0 \in \mathbb{R},\ k = 0$
- While $|\nabla f_k|/|\nabla f_0| > \epsilon$

quNac quasi-Newton action Constrained

- $H_0, x_0 \in \mathbb{R},\ k = 0$
- While $|\nabla f_k|/|\nabla f_0| > \epsilon$
  - If $k = 0$

quNac quasi-Newton action Constrained

- $H_0, x_0 \in \mathbb{R},\ k = 0$
- While $|\nabla f_k|/|\nabla f_0| > \epsilon$
  - **If** $k = 0$
    - $s_0 = -H_0 \nabla f_0$

quNac quasi-Newton action Constrained

- $H_0, x_0 \in \mathbb{R},\ k = 0$
- While $|\nabla f_k| / |\nabla f_0| > \epsilon$
  - **If** $k = 0$
    - $s_0 = -H_0 \nabla f_0$
  - **Else**

quNac quasi-Newton action Constrained

- $H_0, x_0 \in \mathbb{R},\ k = 0$
- While $|\nabla f_k| / |\nabla f_0| > \epsilon$
  - **If** $k = 0$
    - $s_0 = -H_0 \nabla f_0$
  - **Else**
    - $s_0 = d_{CG}$

quNac quasi-Newton action Constrained

- $H_0, x_0 \in \mathbb{R},\ k = 0$
- While $|\nabla f_k|/|\nabla f_0| > \epsilon$
  - **If** $k = 0$
    - $s_0 = -H_0 \nabla f_0$
  - **Else**
    - $s_0 = d_{CG}$
  - Select step-size a_k

quNac quasi-Newton action Constrained

- $H_0, x_0 \in \mathbb{R},\ k = 0$
- While $|\nabla f_k|/|\nabla f_0| > \epsilon$
  - **If** $k = 0$
    - $s_0 = -H_0 \nabla f_0$
  - **Else**
    - $s_0 = d_{CG}$
  - Select step-size a_k
  - $x_{k+1} = x_k + a_k s_k$

quNac quasi-Newton action Constrained

- $H_0, x_0 \in \mathbb{R},\ k = 0$
- While $|\nabla f_k| / |\nabla f_0| > \epsilon$
  - **If** $k = 0$
    - $s_0 = -H_0 \nabla f_0$
  - **Else**
    - $s_0 = d_{CG}$
  - Select step-size a_k
  - $x_{k+1} = x_k + a_k s_k$
  - Calculate d_{CG} by applying CG to preconditioned system

quNac quasi-Newton action Constrained

- $H_0, x_0 \in \mathbb{R},\ k = 0$
- While $|\nabla f_k| / |\nabla f_0| > \epsilon$
  - **If** $k = 0$
    - $s_0 = -H_0 \nabla f_0$
  - **Else**
    - $s_0 = d_{CG}$
  - Select step-size a_k
  - $x_{k+1} = x_k + a_k s_k$
  - Calculate d_{CG} by applying CG to preconditioned system
    - $H_k \nabla^2 f_{k+1} d = -H_k \nabla f_{k+1}.$

quNac quasi-Newton action Constrained

- $H_0, x_0 \in \mathbb{R},\ k = 0$
- While $|\nabla f_k|/|\nabla f_0| > \epsilon$
  - **If** $k = 0$
    - $s_0 = -H_0 \nabla f_0$
  - **Else**
    - $s_0 = d_{CG}$
  - Select step-size a_k
  - $x_{k+1} = x_k + a_k s_k$
  - Calculate d_{CG} by applying CG to preconditioned system
    - $H_k \nabla^2 f_{k+1} d = -H_k \nabla f_{k+1}$.
    - Store conjugate directions in D and store $\nabla^2 f_{k+1} D$.

quNac quasi-Newton action Constrained

- $H_0, x_0 \in \mathbb{R},\ k = 0$
- While $|\nabla f_k|/|\nabla f_0| > \epsilon$
  - **If** $k = 0$
    - $s_0 = -H_0 \nabla f_0$
  - **Else**
    - $s_0 = d_{CG}$
  - Select step-size a_k
  - $x_{k+1} = x_k + a_k s_k$
  - Calculate d_{CG} by applying CG to preconditioned system
    - $H_k \nabla^2 f_{k+1} d = -H_k \nabla f_{k+1}$.
    - Store conjugate directions in D and store $\nabla^2 f_{k+1} D$.
  - $H_{k+1} = H_k + E_k\left(D, \nabla^2 f_{k+1}(D)\right)$

LquNac quasi-Newton action Constrained

- $H_0, x_0 \in \mathbb{R},\ k = 0$
- While $|\nabla f_k|/|\nabla f_0| > \epsilon$
  - **If** $k = 0$
    - $s_0 = -H_0 \nabla f_0$
  - **Else**
    - $s_0 = d_{CG}$
  - Select step-size a_k
  - $x_{k+1} = x_k + a_k s_k$
  - Calculate d_{CG} by applying CG to preconditioned system
    - $H_k \nabla^2 f_{k+1} d = -H_k \nabla f_{k+1}$.
    - Store conjugate directions in D and store $\nabla^2 f_{k+1} D$.
  - $H_{k+1} = H_k + E_k \left(D, \nabla^2 f_{k+1}(D)\right)$

LquNac Limited quasi-Newton action Constrained

- $H_0, x_0 \in \mathbb{R},\ k = 0$
- While $|\nabla f_k|/|\nabla f_0| > \epsilon$
  - **If** $k = 0$
    - $s_0 = -H_0 \nabla f_0$
  - **Else**
    - $s_0 = d_{CG}$
  - Select step-size a_k
  - $x_{k+1} = x_k + a_k s_k$
  - Calculate d_{CG} by applying CG to preconditioned system
    - $OpH_k \nabla^2 f_{k+1} d = -OpH_k \nabla f_{k+1}$.
    - Store conjugate directions in D and store $\nabla^2 f_{k+1} D$.
  - $OpH_{k+1} = H_0^k + E_k \left(D, \nabla^2 f_{k+1}(D)\right)$

Logistic L2 Regression tests:

$\min_w L_w(y,X) + \|w\|_2^2$

$$L_w(y,X) = \sum_{i=1}^{m} \ln\left(1 + \exp(-y_i \left\langle x^i, w \right\rangle)\right).$$

quNac	vs	BFGS
41		3
quNac	vs	Newton_CG
31		12
LquNac	vs	LBFGS
27		17
LquNac	vs	Newton_CG
14		29

Table: # fastest runs on 44 binary classifications problems from LibSVM

Logistic pseudo-Huber Regression tests:

$$\min_{w} L_w(y, X) + R_\mu(w) := \mu \sum_{i=1}^{n} \left(\sqrt{1 + x_i^2/\mu^2} - 1 \right).$$ [4]

quNac	vs	BFGS
32		10
quNac	vs	Newton_CG
37		4
LquNac	vs	LBFGS
18		25
LquNac	vs	Newton_CG
24		16

Table: # fastest runs on 44 binary classifications problems from LibSVM

[4] Fountoulakis, K., & Gondzio, J. (2013). A Second-Order Method for Strongly Convex l1-regularization Problems.

Conclusion

- Framework for approximating a changing inverse Hessian.
- Has good properties: **Hereditary** and **Descent**.
- Variable amount of curvature information at each iteration (depends on CG error).
- Developed Newton-CG Preconditioner for smooth functions.

In the works:

- Full and limited memory implementations for non-convex unconstrained.
- How does extra flexibility help mesh into globalization strategies: Interior point, Trust region, Sequential Quadratic?
- Connections to other problems: Matrix completion?

Conclusion

- Framework for approximating a changing inverse Hessian.
- Has good properties: **Hereditary** and **Descent**.
- Variable amount of curvature information at each iteration (depends on CG error).
- Developed Newton-CG Preconditioner for smooth functions.

In the works:

- Full and limited memory implementations for non-convex unconstrained.
- How does extra flexibility help mesh into globalization strategies: Interior point, Trust region, Sequential Quadratic?
- Connections to other problems: Matrix completion?

References

Gower, R. M., (2014).
New quasi-Newton family through action constraints (*in preparation*)

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.

Shanno, D. F. (1971).
Conditioning of Quasi-Newton Methods for Function Minimization.
Mathematics of Computation, 24(111), 647656.

Underlying Matrix Optimization Problem

$$\begin{aligned} &\min_E \|E\|^2_{Frobenius(W)} \\ \text{s.t.} \quad & ED = RD \\ & E = E^T \end{aligned}$$

Which has a low rank$-3p$ solution.

$$E = W^{-1}\text{proj}_D^{W^{-1}} R \left(I - \text{proj}_D^{W^{-1}} W^{-1} \right) + R\text{proj}_D^{W^{-1}} W^{-1}.$$

This is a matrix completion problem where one knows the desired matrix is symmetric and can only observe its action on a small subspace.