Conjugate gradient training algorithm

So far:

- Heuristic improvements to gradient descent (momentum)
- Steepest descent training algorithm
- Can we do better?

Next: *Conjugate gradient training algorithm*

- Overview
- Derivation
- Examples

Steepest descent algorithm

Definitions:

- $\mathbf{w}_j$ = weight vector at step j .
- $\mathbf{g}_j = \nabla E[\mathbf{w}_j]$ = gradient at step j .
- $\mathbf{d}_j$ = search direction at step j .

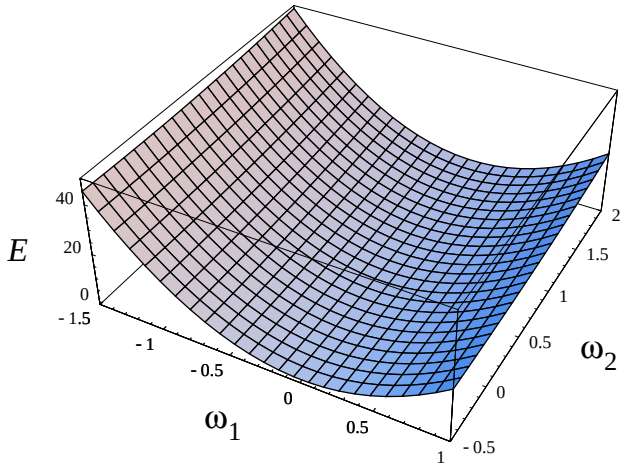
Steepest descent algorithm

1. Choose an initial weight vector $\mathbf{w}_1$ and let $\mathbf{d}_1 = -\mathbf{g}_1$.
2. Perform line minimization along $\mathbf{d}_j$, $j \geq 1$, such that:

$$E(\mathbf{w}_j + \eta^* \mathbf{d}_j) \leq E(\mathbf{w}_j + \eta \mathbf{d}_j) , \ \forall \eta .$$

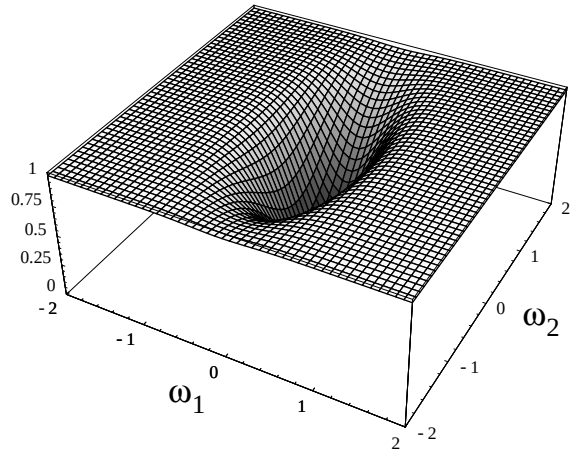
3. Let $\mathbf{w}_{j+1} = \mathbf{w}_j + \eta^* \mathbf{d}_j$.
4. Evaluate $\mathbf{g}_{j+1}$.
5. Let $\mathbf{d}_{j+1} = -\mathbf{g}_{j+1}$.
6. Let $j = j + 1$ and go to step 2.

Remember previous examples



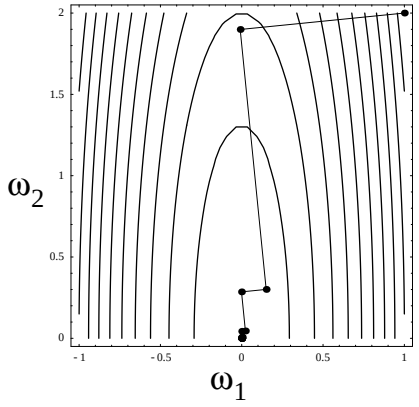
$$E = 20\omega_1^2 + \omega_2^2$$

Remember previous examples



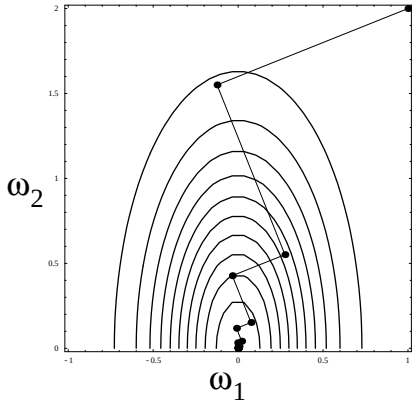
$$E(\omega_1, \omega_2) = 1 - \exp(-5\omega_1^2 - \omega_2^2)$$

Steepest descent algorithm examples



steepest descent
15 steps to convergence

quadratic



steepest descent
24 steps to convergence

non-quadratic

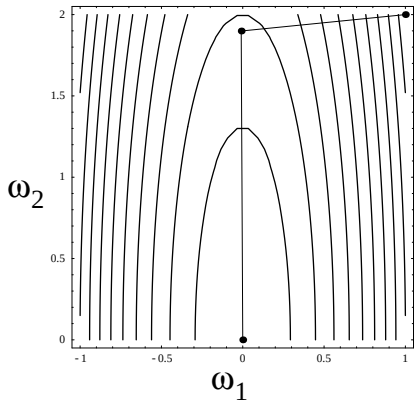
Conjugate gradient algorithm (a sneak peek)

1. Choose an initial weight vector $\mathbf{w}_1$ and let $\mathbf{d}_1 = -\mathbf{g}_1$.
2. Perform a line minimization along $\mathbf{d}_j$, such that:

   $E(\mathbf{w}_j + \eta^* \mathbf{d}_j) \leq E(\mathbf{w}_j + \eta \mathbf{d}_j)$, $\forall \eta$.
3. Let $\mathbf{w}_{j+1} = \mathbf{w}_j + \eta^* \mathbf{d}_j$.
4. Evaluate $\mathbf{g}_{j+1}$.
5. Let $\mathbf{d}_{j+1} = -\mathbf{g}_{j+1} + \beta_j \mathbf{d}_j$ where,

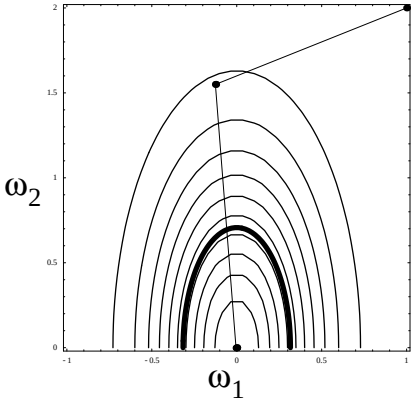
   $$\beta_j = \frac{\mathbf{g}_{j+1}^T (\mathbf{g}_{j+1} - \mathbf{g}_j)}{\mathbf{g}_j^T \mathbf{g}_j}$$
6. Let $j = j + 1$ and go to step 2.

Conjugate gradient algorithm (sneak peak)



conjugate gradient
2 steps to convergence

quadratic



conjugate gradient
5 steps to convergence

non-quadratic

SD *vs.* CG

Key difference: new search direction

- Very little additional computation (over steepest descent).
- No more oscillation back and forth.

Key question:

- Why/How does this improve things?

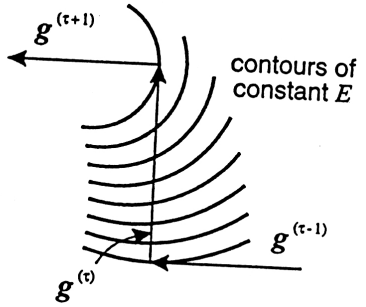
Knowledge of local quadratic properties of error surface...

Conjugate gradients: a first look

In steepest descent:

$$\mathbf{g}[\mathbf{w}(t+1)]^T\mathbf{d}(t) = 0$$

(What does this mean?)

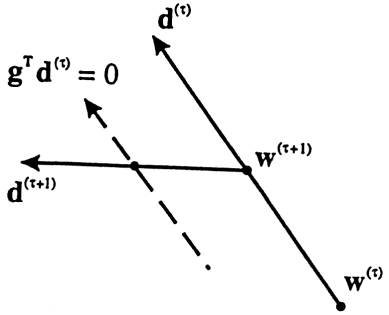


Conjugate gradients: a first look

Non-interfering directions:

$$\mathbf{g}[\mathbf{w}(t+1) + \eta\mathbf{d}(t+1)]^T\mathbf{d}(t) = 0\,,\ \eta \geq 0$$

(What the #$@!# does this mean?)



How do we achieve non-interfering directions?

FACT:

$$\mathbf{g}[\mathbf{w}(t+1) + \eta\mathbf{d}(t+1)]^T\mathbf{d}(t) = 0\,,\ \eta \geq 0$$

implies

$$\mathbf{d}(t+1)^T\mathbf{H}\mathbf{d}(t) = 0 \quad \textit{(H-orthogonality, conjugacy)}$$

Hmmm ... need to pay attention to 2nd-order properties of error surface...

$$E(\mathbf{w}) = E_0 + \mathbf{b}^T\mathbf{w} + \frac{1}{2}\mathbf{w}^T\mathbf{H}\mathbf{w}$$

Show H-orthogonality requirement

Approximate $\mathbf{g}(\mathbf{w})$ by 1st-order Taylor approximation about $\mathbf{w}_0$:

$$\mathbf{g}(\mathbf{w})^T \approx \mathbf{g}(\mathbf{w}_0)^T + (\mathbf{w} - \mathbf{w}_0)^T \nabla\{\mathbf{g}(\mathbf{w}_0)\}$$

Let $\mathbf{w}_0 = \mathbf{w}(t+1)$:

$$\mathbf{g}(\mathbf{w})^T = \mathbf{g}[\mathbf{w}(t+1)]^T + [\mathbf{w} - \mathbf{w}(t+1)]^T \nabla\{\mathbf{g}[\mathbf{w}(t+1)]\}$$

Evaluate $\mathbf{g}(\mathbf{w})$ at $\mathbf{w} = \mathbf{w}(t+1) + \eta\mathbf{d}(t+1)$:

$$\mathbf{g}[\mathbf{w}(t+1) + \eta\mathbf{d}(t+1)]^T = \mathbf{g}[\mathbf{w}(t+1)]^T + \eta\mathbf{d}(t+1)^T\mathbf{H}$$

Show H-orthogonality requirement

$$\mathbf{g}[\mathbf{w}(t+1) + \eta\mathbf{d}(t+1)]^T = \mathbf{g}[\mathbf{w}(t+1)]^T + \eta\mathbf{d}(t+1)^T\mathbf{H}$$

Post-multiply by $\mathbf{d}(t)$:

- Left-hand side:

$$\mathbf{g}[\mathbf{w}(t+1) + \eta\mathbf{d}(t+1)]^T\mathbf{d}(t) = 0 \text{ (assumption)}$$

- Right-hand side:

$$\mathbf{g}[\mathbf{w}(t+1)]^T\mathbf{d}(t) + \eta\mathbf{d}(t+1)^T\mathbf{H}\mathbf{d}(t) = 0$$

$$\eta\mathbf{d}(t+1)^T\mathbf{H}\mathbf{d}(t) = 0$$

$$\mathbf{d}(t+1)^T\mathbf{H}\mathbf{d}(t) = 0 \text{ (implication)}$$

How do we achieve non-interfering directions?

Proven fact:

$$\mathbf{g}[\mathbf{w}(t+1) + \eta\mathbf{d}(t+1)]^T\mathbf{d}(t) = 0\,,\ \eta \geq 0$$

implies

$$\mathbf{d}(t+1)^T\mathbf{H}\mathbf{d}(t) = 0 \text{ } \textit{(H-orthogonality)}$$

Key: need to construct consecutive search directions $\mathbf{d}$ that are conjugate ($\mathbf{H}$-orthogonal)!

(Side note: What is the implicit assumption of SD?)

Derivation of conjugate gradient algorithm

- **Local quadratic assumption:**

$$E(\mathbf{w}) = E_0 + \mathbf{b}^T\mathbf{w} + \frac{1}{2}\mathbf{w}^T\mathbf{H}\mathbf{w}$$

- **Assume:**
  - W mutually conjugate vectors $\mathbf{d}_i$.
  - Initial weight vector $\mathbf{w}_1$.

Question: How to converge to $\mathbf{w}^*$ in (at most) W steps?

Step-wise optimization

$$(\mathbf{w}^* - \mathbf{w}_1) = \sum_{i=1}^{W} \alpha_i \mathbf{d}_i \quad \textit{(why can I do this?)}$$

$$\mathbf{w}^* = \mathbf{w}_1 + \sum_{i=1}^{W} \alpha_i \mathbf{d}_i$$

$$\mathbf{w}_j \equiv \mathbf{w}_1 + \sum_{i=1}^{j-1} \alpha_i \mathbf{d}_i$$

$$\mathbf{w}_{j+1} = \mathbf{w}_j + \alpha_j \mathbf{d}_j$$

Linear independence of conjugate directions

Theorem: **For a positive-definite square matrix $\mathbf{H}$, $\mathbf{H}$-orthogonal vectors $\{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_k\}$ are *linearly independent.***

Proof: **Linear independence:**

$$\alpha_1 \mathbf{d}_1 + \alpha_2 \mathbf{d}_2 + \ldots + \alpha_k \mathbf{d}_k = \mathbf{0} \text{ iff. } \alpha_i = 0, \forall i .$$

Linear independence of conjugate directions

Linear independence:

$$\alpha_1 \mathbf{d}_1 + \alpha_2 \mathbf{d}_2 + \ldots + \alpha_k \mathbf{d}_k = \mathbf{0} \text{ iff. } \alpha_i = 0, \forall i .$$

Pre-multiply by $\mathbf{d}_i^T \mathbf{H}$:

$$\alpha_1 \mathbf{d}_i^T \mathbf{H} \mathbf{d}_1 + \alpha_2 \mathbf{d}_i^T \mathbf{H} \mathbf{d}_2 + \ldots + \alpha_k \mathbf{d}_i^T \mathbf{H} \mathbf{d}_k = \mathbf{d}_i^T \mathbf{H} \mathbf{0}$$

$$\alpha_1 \mathbf{d}_i^T \mathbf{H} \mathbf{d}_1 + \alpha_2 \mathbf{d}_i^T \mathbf{H} \mathbf{d}_2 + \ldots + \alpha_k \mathbf{d}_i^T \mathbf{H} \mathbf{d}_k = 0$$

Note (by assumption): $\mathbf{d}_i^T \mathbf{H} \mathbf{d}_j = 0, \forall i \neq j .$

Linear independence of conjugate directions

$$\alpha_1 \mathbf{d}_i^T \mathbf{H} \mathbf{d}_1 + \alpha_2 \mathbf{d}_i^T \mathbf{H} \mathbf{d}_2 + \ldots + \alpha_k \mathbf{d}_i^T \mathbf{H} \mathbf{d}_k = 0$$

reduces to:

$$\alpha_i \mathbf{d}_i^T \mathbf{H} \mathbf{d}_i = 0$$

However:

$$\mathbf{d}_i^T \mathbf{H} \mathbf{d}_i > 0 \quad \textit{(by assumption)}$$

Therefore:

$$\alpha_i = 0, \; i \in \{1, 2, \ldots, k\} \; \square$$

Linear independence of conjugate directions

From linear independence:

- $\mathbf{H}$-orthogonal vectors $\mathbf{d}_i$ form a complete basis set.
- Any vector $\mathbf{v}$ can be expressed as:

$$\mathbf{v} = \sum_{i=1}^{W} \alpha_i \mathbf{d}_i$$

So, why did we need this result?

Step-wise optimization

$$(\mathbf{w}^* - \mathbf{w}_1) = \sum_{i=1}^{W} \alpha_i \mathbf{d}_i \quad \textit{(Ah-ha!)}$$

$$\mathbf{w}^* = \mathbf{w}_1 + \sum_{i=1}^{W} \alpha_i \mathbf{d}_i$$

$$\mathbf{w}_j \equiv \mathbf{w}_1 + \sum_{i=1}^{j-1} \alpha_i \mathbf{d}_i$$

$$\mathbf{w}_{j+1} = \mathbf{w}_j + \alpha_j \mathbf{d}_j$$

So where are we now?

On locally quadratic surface, can converge to minimum in, at most, W steps using:

$$\mathbf{w}_{j+1} = \mathbf{w}_j + \alpha_j \mathbf{d}_j, \; j = \{1, 2, \ldots, W\}.$$

Big Questions:

- How to choose step size α_j?
- How to construct conjugate directions $\mathbf{d}_j$?
- How can we do everything without computing $\mathbf{H}$?

Computing the correct step size α_j

Given: a set of W conjugate vectors $\mathbf{d}_i$.

$$(\mathbf{w}^* - \mathbf{w}_1) = \sum_{i=1}^{W} \alpha_i \mathbf{d}_i$$

Pre-multiply by $\mathbf{d}_j^T \mathbf{H}$:

$$\mathbf{d}_j^T \mathbf{H}(\mathbf{w}^* - \mathbf{w}_1) = \mathbf{d}_j^T \mathbf{H} \left(\sum_{i=1}^{W} \alpha_i \mathbf{d}_i \right)$$

$$\mathbf{d}_j^T \mathbf{H}(\mathbf{w}^* - \mathbf{w}_1) = \sum_{i=1}^{W} \alpha_i \mathbf{d}_j^T \mathbf{H} \mathbf{d}_i$$

Computing the correct step size α_j

$$\mathbf{d}_j^T\mathbf{H}(\mathbf{w}^* - \mathbf{w}_1) = \sum_{i=1}^{W} \alpha_i \mathbf{d}_j^T\mathbf{H}\mathbf{d}_i$$

By $\mathbf{H}$-orthogonality (conjugacy assumed):

$$\mathbf{d}_j^T\mathbf{H}(\mathbf{w}^* - \mathbf{w}_1) = \alpha_j \mathbf{d}_j^T\mathbf{H}\mathbf{d}_j \quad \textit{(why?)}$$

Computing the correct step size α_j

Also:

$$E(\mathbf{w}) = E_0 + \mathbf{b}^T\mathbf{w} + \frac{1}{2}\mathbf{w}^T\mathbf{H}\mathbf{w}$$

$$\mathbf{g}(\mathbf{w}) = \mathbf{b} + \mathbf{H}\mathbf{w}$$

At minimum:

$$\mathbf{g}(\mathbf{w}^*) = 0$$

$$\mathbf{b} + \mathbf{H}\mathbf{w}^* = 0$$

$$\mathbf{H}\mathbf{w}^* = -\mathbf{b}$$

Computing the correct step size α_j

$$\mathbf{d}_j^T\mathbf{H}(\mathbf{w}^* - \mathbf{w}_1) = \alpha_j \mathbf{d}_j^T\mathbf{H}\mathbf{d}_j$$

$$\mathbf{H}\mathbf{w}^* = -\mathbf{b}$$

So:

$$\mathbf{d}_j^T(-\mathbf{b} - \mathbf{H}\mathbf{w}_1) = \alpha_j \mathbf{d}_j^T\mathbf{H}\mathbf{d}_j$$

$$-\mathbf{d}_j^T(\mathbf{b} + \mathbf{H}\mathbf{w}_1) = \alpha_j \mathbf{d}_j^T\mathbf{H}\mathbf{d}_j$$

$$\alpha_j = \frac{-\mathbf{d}_j^T(\mathbf{b} + \mathbf{H}\mathbf{w}_1)}{\mathbf{d}_j^T\mathbf{H}\mathbf{d}_j} \quad \textit{(what's the problem?)}$$

Computing the correct step size α_j

$$\alpha_j = \frac{-\mathbf{d}_j^T(\mathbf{b} + \mathbf{H}\mathbf{w}_1)}{\mathbf{d}_j^T\mathbf{H}\mathbf{d}_j} \qquad \mathbf{w}_j = \mathbf{w}_1 + \sum_{i=1}^{j-1} \alpha_i \mathbf{d}_i$$

Pre-multiply by $\mathbf{d}_j^T\mathbf{H}$:

$$\mathbf{d}_j^T\mathbf{H}\mathbf{w}_j = \mathbf{d}_j^T\mathbf{H}\mathbf{w}_1 + \sum_{i=1}^{j-1} \alpha_i \mathbf{d}_j^T\mathbf{H}\mathbf{d}_i \,.$$

$$\mathbf{d}_j^T\mathbf{H}\mathbf{w}_j = \mathbf{d}_j^T\mathbf{H}\mathbf{w}_1 \,.$$

$$\alpha_j = \frac{-\mathbf{d}_j^T(\mathbf{b} + \mathbf{H}\mathbf{w}_j)}{\mathbf{d}_j^T\mathbf{H}\mathbf{d}_j}$$

Computing the correct step size α_j

$$\alpha_j = \frac{-\mathbf{d}_j^T(\mathbf{b} + \mathbf{H}\mathbf{w}_j)}{\mathbf{d}_j^T\mathbf{H}\mathbf{d}_j}$$

$$\mathbf{g}_j = \mathbf{b} + \mathbf{H}\mathbf{w}_j$$

So:

$$\alpha_j = \frac{-\mathbf{d}_j^T\mathbf{g}_j}{\mathbf{d}_j^T\mathbf{H}\mathbf{d}_j}$$ *(woo-hoo!)*

Important consequence

Theorem: **Assuming a W-dimensional quadratic error surface,**

$$E(\mathbf{w}) = E_0 + \mathbf{b}^T\mathbf{w} + \frac{1}{2}\mathbf{w}^T\mathbf{H}\mathbf{w}$$

and $\mathbf{H}$-orthogonal vectors $\mathbf{d}_i$, $i \in \{1, 2, \ldots, W\}$:

$$\mathbf{w}_{j+1} = \mathbf{w}_j + \alpha_j\mathbf{d}_j$$

$$\alpha_j = \frac{-\mathbf{d}_j^T\mathbf{g}_j}{\mathbf{d}_j^T\mathbf{H}\mathbf{d}_j}$$

will converge in at most W steps to the minimum $\mathbf{w}^*$ (for what error surface?).

Why is this so?

FACT:

$$\mathbf{d}_k^T\mathbf{g}_j = 0\ ,\ \forall k < j$$

How is this important?

How is this different from steepest descent?

Let's show that this is true...

Orthogonality of gradient to previous search directions

$$\mathbf{g}_j = \mathbf{b} + \mathbf{H}\mathbf{w}_j$$

$$\mathbf{w}_{j+1} = \mathbf{w}_j + \alpha_j\mathbf{d}_j$$

So:

$$\mathbf{g}_{j+1} - \mathbf{g}_j = \mathbf{H}(\mathbf{w}_{j+1} - \mathbf{w}_j)$$

$$(\mathbf{w}_{j+1} - \mathbf{w}_j) = \alpha_j\mathbf{d}_j$$

$$\mathbf{g}_{j+1} - \mathbf{g}_j = \alpha_j\mathbf{H}\mathbf{d}_j$$

Orthogonality of gradient to previous search directions

$$\mathbf{g}_{j+1} - \mathbf{g}_j = \alpha_j \mathbf{H}\mathbf{d}_j$$

$$\alpha_j = \frac{-\mathbf{d}_j^T \mathbf{g}_j}{\mathbf{d}_j^T \mathbf{H}\mathbf{d}_j}$$

Pre-multiply by $\mathbf{d}_j^T$:

$$\mathbf{d}_j^T(\mathbf{g}_{j+1} - \mathbf{g}_j) = \alpha_j \mathbf{d}_j^T \mathbf{H}\mathbf{d}_j$$

$$\mathbf{d}_j^T(\mathbf{g}_{j+1} - \mathbf{g}_j) = \left(\frac{-\mathbf{d}_j^T \mathbf{g}_j}{\mathbf{d}_j^T \mathbf{H}\mathbf{d}_j}\right)\mathbf{d}_j^T \mathbf{H}\mathbf{d}_j$$

Orthogonality of gradient to previous search directions

$$\mathbf{d}_j^T(\mathbf{g}_{j+1} - \mathbf{g}_j) = \left(\frac{-\mathbf{d}_j^T \mathbf{g}_j}{\mathbf{d}_j^T \mathbf{H}\mathbf{d}_j}\right)\mathbf{d}_j^T \mathbf{H}\mathbf{d}_j$$

So:

$$\mathbf{d}_j^T(\mathbf{g}_{j+1} - \mathbf{g}_j) = -\mathbf{d}_j^T \mathbf{g}_j$$

$$\mathbf{d}_j^T \mathbf{g}_{j+1} - \mathbf{d}_j^T \mathbf{g}_j = -\mathbf{d}_j^T \mathbf{g}_j$$

$$\mathbf{d}_j^T \mathbf{g}_{j+1} = 0$$

$\mathbf{d}_k^T \mathbf{g}_j = 0$, $k = j-1$ *(need to show for all $k < j$)*

Orthogonality of gradient to previous search directions

$$\mathbf{g}_{j+1} - \mathbf{g}_j = \alpha_j \mathbf{H}\mathbf{d}_j$$

Pre-multiply by $\mathbf{d}_k^T$:

$$\mathbf{d}_k^T(\mathbf{g}_{j+1} - \mathbf{g}_j) = \alpha_j \mathbf{d}_k^T \mathbf{H}\mathbf{d}_j \ .$$

$\mathbf{d}_k^T(\mathbf{g}_{j+1} - \mathbf{g}_j) = 0$ *(why?)*

$\mathbf{d}_k^T \mathbf{g}_{j+1} = \mathbf{d}_k^T \mathbf{g}_j$, $k < j$

Orthogonality of gradient to previous search directions

$$\mathbf{d}_j^T \mathbf{g}_{j+1} = 0$$

$$\mathbf{d}_k^T \mathbf{g}_{j+1} = \mathbf{d}_k^T \mathbf{g}_j \ , \ k < j$$

- **By induction:** $\mathbf{d}_k^T \mathbf{g}_j = 0$, $\forall k < j$
- **For example:**

$$\mathbf{d}_{j-1}^T \mathbf{g}_{j+1} = \mathbf{d}_{j-1}^T \mathbf{g}_j$$

$$\mathbf{d}_{j-1}^T \mathbf{g}_j = 0$$

$$\mathbf{d}_{j-1}^T \mathbf{g}_{j+1} = 0$$

So where are we now?

On locally quadratic surface, can converge to minimum in, at most, W steps using:

$$\mathbf{w}_{j+1} = \mathbf{w}_j + \alpha_j \mathbf{d}_j, \; j = \{1, 2, \ldots, W\}.$$

$$\alpha_j = \frac{-\mathbf{d}_j^T \mathbf{g}_j}{\mathbf{d}_j^T \mathbf{H} \mathbf{d}_j}$$

Remaining Big Questions:

- How to construct conjugate directions $\mathbf{d}_j$?
- How can we do everything without computing $\mathbf{H}$?

Constructing conjugate directions $\mathbf{d}_j$

Theorem: **Let $\mathbf{d}_j$ be defined as follows:**

1. Let $\mathbf{d}_1 = -\mathbf{g}_1$.

2. Let,

$$\mathbf{d}_{j+1} = -\mathbf{g}_{j+1} + \beta_j \mathbf{d}_j, \; j \geq 1,$$

$$\beta_j = \frac{\mathbf{g}_{j+1}^T \mathbf{H} \mathbf{d}_j}{\mathbf{d}_j^T \mathbf{H} \mathbf{d}_j}$$

This construction generates W mutually $\mathbf{H}$-orthogonal vectors, such that:

$$\mathbf{d}_j^T \mathbf{H} \mathbf{d}_i = 0, \; \forall i \neq j.$$

Constructing conjugate directions $\mathbf{d}_j$: Part I

First goal: Show that,

$$\mathbf{d}_{j+1}^T \mathbf{H} \mathbf{d}_j = 0$$

Begin with:

$$\mathbf{d}_{j+1} = -\mathbf{g}_{j+1} + \beta_j \mathbf{d}_j$$

Pre-multiply by $\mathbf{d}_j^T \mathbf{H}$:

$$\mathbf{d}_j^T \mathbf{H} \mathbf{d}_{j+1} = -\mathbf{d}_j^T \mathbf{H} \mathbf{g}_{j+1} + \beta_j \mathbf{d}_j^T \mathbf{H} \mathbf{d}_j$$

Constructing conjugate directions $\mathbf{d}_j$: Part I

$$\mathbf{d}_j^T \mathbf{H} \mathbf{d}_{j+1} = -\mathbf{d}_j^T \mathbf{H} \mathbf{g}_{j+1} + \beta_j \mathbf{d}_j^T \mathbf{H} \mathbf{d}_j$$

$$\beta_j = \frac{\mathbf{g}_{j+1}^T \mathbf{H} \mathbf{d}_j}{\mathbf{d}_j^T \mathbf{H} \mathbf{d}_j}$$

So:

$$\mathbf{d}_j^T \mathbf{H} \mathbf{d}_{j+1} = -\mathbf{d}_j^T \mathbf{H} \mathbf{g}_{j+1} + \left(\frac{\mathbf{g}_{j+1}^T \mathbf{H} \mathbf{d}_j}{\mathbf{d}_j^T \mathbf{H} \mathbf{d}_j} \right) \mathbf{d}_j^T \mathbf{H} \mathbf{d}_j$$

$$\mathbf{d}_j^T \mathbf{H} \mathbf{d}_{j+1} = -\mathbf{d}_j^T \mathbf{H} \mathbf{g}_{j+1} + \mathbf{g}_{j+1}^T \mathbf{H} \mathbf{d}_j = 0$$

$$\mathbf{d}_{j+1}^T \mathbf{H} \mathbf{d}_j = 0$$

Constructing conjugate directions $\mathbf{d}_j$: Part II

Second goal, show that:

$$\mathbf{d}_j^T \mathbf{H} \mathbf{d}_i = 0 \quad \Rightarrow \quad \mathbf{d}_{j+1}^T \mathbf{H} \mathbf{d}_i = 0 \,, \; \forall i < j$$

Begin with:

$$\mathbf{d}_{j+1} = -\mathbf{g}_{j+1} + \beta_j \mathbf{d}_j$$

Transpose and post-multiply by $\mathbf{H}\mathbf{d}_i$, $i < j$:

$$\mathbf{d}_{j+1}^T \mathbf{H} \mathbf{d}_i = -\mathbf{g}_{j+1}^T \mathbf{H} \mathbf{d}_i + \beta_j \mathbf{d}_j^T \mathbf{H} \mathbf{d}_i$$

$$\mathbf{d}_{j+1}^T \mathbf{H} \mathbf{d}_i = -\mathbf{g}_{j+1}^T \mathbf{H} \mathbf{d}_i \quad \textit{(by assumption)}$$

Constructing conjugate directions $\mathbf{d}_j$: Part II

$$\mathbf{w}_{j+1} = \mathbf{w}_j + \alpha_j \mathbf{d}_j$$

So:

$$\mathbf{w}_{j+1} - \mathbf{w}_j = \alpha_j \mathbf{d}_j$$

$$\mathbf{H}(\mathbf{w}_{j+1} - \mathbf{w}_j) = \alpha_j \mathbf{H} \mathbf{d}_j$$

Remember that:

$$\mathbf{g}_j = \mathbf{H}\mathbf{w}_j + \mathbf{b}$$

$$\mathbf{H}(\mathbf{w}_{j+1} - \mathbf{w}_j) = \mathbf{g}_{j+1} - \mathbf{g}_j$$

$$\alpha_j \mathbf{H} \mathbf{d}_j = \mathbf{g}_{j+1} - \mathbf{g}_j \quad \textit{(why?)}$$

Constructing conjugate directions $\mathbf{d}_j$: Part II

$$\alpha_j \mathbf{H} \mathbf{d}_j = \mathbf{g}_{j+1} - \mathbf{g}_j$$

$$\mathbf{H} \mathbf{d}_j = \frac{1}{\alpha_j}(\mathbf{g}_{j+1} - \mathbf{g}_j)$$

$$\mathbf{d}_{j+1}^T \mathbf{H} \mathbf{d}_i = -\mathbf{g}_{j+1}^T \mathbf{H} \mathbf{d}_i \quad \textit{(from before)}$$

So:

$$\mathbf{d}_{j+1}^T \mathbf{H} \mathbf{d}_i = -\frac{1}{\alpha_i} \mathbf{g}_{j+1}^T (\mathbf{g}_{i+1} - \mathbf{g}_i)$$

Constructing conjugate directions $\mathbf{d}_j$: Part II

$$\mathbf{d}_{j+1}^T \mathbf{H} \mathbf{d}_i = -\frac{1}{\alpha_i} \mathbf{g}_{j+1}^T (\mathbf{g}_{i+1} - \mathbf{g}_i)$$

So:

$$\mathbf{d}_{j+1}^T \mathbf{H} \mathbf{d}_i = -\frac{1}{\alpha_i} (\mathbf{g}_{j+1}^T \mathbf{g}_{i+1} - \mathbf{g}_{j+1}^T \mathbf{g}_i)$$

$$\mathbf{d}_{j+1}^T \mathbf{H} \mathbf{d}_i = -\frac{1}{\alpha_i} (\mathbf{g}_{i+1}^T \mathbf{g}_{j+1} - \mathbf{g}_i^T \mathbf{g}_{j+1}) \,, \; i < j$$

Constructing conjugate directions $\mathbf{d}_j$: Part II

$$\mathbf{d}_{j+1}^T \mathbf{H} \mathbf{d}_i = -\frac{1}{\alpha_i}(\mathbf{g}_{i+1}^T \mathbf{g}_{j+1} - \mathbf{g}_i^T \mathbf{g}_{j+1}) \text{ , } i < j$$

Now need to show that:

$$\mathbf{g}_k^T \mathbf{g}_j = 0 \text{ , } \forall k < j$$

so that

$$\mathbf{d}_j^T \mathbf{H} \mathbf{d}_i = 0 \quad \Rightarrow \quad \mathbf{d}_{j+1}^T \mathbf{H} \mathbf{d}_i = 0 \text{ , } \forall i < j$$

is proven.

Constructing conjugate directions $\mathbf{d}_j$: Part II

From construction:

$$\mathbf{d}_1 = -\mathbf{g}_1$$

$$\mathbf{d}_{j+1} = -\mathbf{g}_{j+1} + \beta_j \mathbf{d}_j \text{ , } j \geq 1$$

we see that:

$$\mathbf{d}_k = -\mathbf{g}_k + \sum_{l=1}^{k-1} \gamma_l \mathbf{g}_l \quad \textit{(why?)}$$

Constructing conjugate directions $\mathbf{d}_j$: Part II

For example:

$$\mathbf{d}_1 = -\mathbf{g}_1$$

$$\mathbf{d}_2 = -\mathbf{g}_2 + \beta_1 \mathbf{d}_1 = -\mathbf{g}_2 - \beta_1 \mathbf{g}_1$$

$$\mathbf{d}_3 = -\mathbf{g}_3 + \beta_2 \mathbf{d}_2 = -\mathbf{g}_3 - \beta_2 \mathbf{g}_2 - \beta_2 \beta_1 \mathbf{g}_1$$

Constructing conjugate directions $\mathbf{d}_j$: Part II

$$\mathbf{d}_k = -\mathbf{g}_k + \sum_{l=1}^{k-1} \gamma_l \mathbf{g}_l$$

Transpose and post-multiply by $\mathbf{g}_j$, $k < j$:

$$\mathbf{d}_k^T \mathbf{g}_j = -\mathbf{g}_k^T \mathbf{g}_j + \sum_{l=1}^{k-1} \gamma_l \mathbf{g}_l^T \mathbf{g}_k \text{ ,}$$

$$0 = -\mathbf{g}_k^T \mathbf{g}_j + \sum_{l=1}^{k-1} \gamma_l \mathbf{g}_l^T \mathbf{g}_k \quad \textit{(why?)}$$

$$\mathbf{d}_k^T \mathbf{g}_j = 0 \text{ , } \forall k < j \quad \textit{(because...)}$$

Constructing conjugate directions $\mathbf{d}_j$: Part II

$$0 = -\mathbf{g}_k^T\mathbf{g}_j + \sum_{l=1}^{k-1} \gamma_l \mathbf{g}_l^T\mathbf{g}_k \qquad \mathbf{d}_k^T\mathbf{g}_j = 0, \ \forall k < j$$

So:

$$\mathbf{g}_k^T\mathbf{g}_j = \sum_{l=1}^{k-1} \gamma_l \mathbf{g}_l^T\mathbf{g}_k$$

Since $\mathbf{d}_1 = -\mathbf{g}_1$ **:**

$$\mathbf{d}_1^T\mathbf{g}_j = -\mathbf{g}_1^T\mathbf{g}_j = 0 \ , \ j > 1 \ \textit{(why?)}$$

$$\mathbf{g}_1^T\mathbf{g}_j = 0 \ , \ j > 1$$

Constructing conjugate directions $\mathbf{d}_j$: Part II

$$\mathbf{g}_k^T\mathbf{g}_j = \sum_{l=1}^{k-1} \gamma_l \mathbf{g}_l^T\mathbf{g}_k \ , \ k < j \qquad \mathbf{g}_1^T\mathbf{g}_j = 0 \ , \ j > 1$$

By induction:

$$\mathbf{g}_1^T\mathbf{g}_2 = 0 \ \textit{(why?)}$$

$$\mathbf{g}_2^T\mathbf{g}_3 = \gamma_1 \mathbf{g}_1^T\mathbf{g}_3 = 0 \ \textit{(why?)}$$

$$\mathbf{g}_1^T\mathbf{g}_4 = 0 \ \textit{(why?)}$$

$$\mathbf{g}_2^T\mathbf{g}_4 = \gamma_1 \mathbf{g}_1^T\mathbf{g}_4 = 0 \ \textit{(why?)}$$

$$\mathbf{g}_3^T\mathbf{g}_4 = \gamma_1 \mathbf{g}_1^T\mathbf{g}_4 + \gamma_2 \mathbf{g}_2^T\mathbf{g}_4 = 0 \ \textit{(why?)}$$

Constructing conjugate directions $\mathbf{d}_j$: Part II

Thus:

$$\mathbf{g}_k^T\mathbf{g}_j = 0 \ , \ \forall k < j$$

So:

$$\mathbf{d}_{j+1}^T\mathbf{H}\mathbf{d}_i = -\frac{1}{\alpha_i}(\mathbf{g}_{i+1}^T\mathbf{g}_{j+1} - \mathbf{g}_i^T\mathbf{g}_{j+1}) = 0 \ , \ i < j$$

$$\mathbf{d}_j^T\mathbf{H}\mathbf{d}_i = 0 \quad \Rightarrow \quad \mathbf{d}_{j+1}^T\mathbf{H}\mathbf{d}_i = 0 \ , \ \forall i < j$$

Constructing conjugate directions $\mathbf{d}_j$

Where we are:

$$\mathbf{d}_{j+1}^T\mathbf{H}\mathbf{d}_j = 0 \ \textit{(Part I)}$$

$$\mathbf{d}_j^T\mathbf{H}\mathbf{d}_i = 0 \quad \Rightarrow \quad \mathbf{d}_{j+1}^T\mathbf{H}\mathbf{d}_i = 0 \ , \ \forall i < j \ \textit{(Part II)}$$

Does this show what we want?

$$\mathbf{d}_j^T\mathbf{H}\mathbf{d}_i = 0 \ , \ \forall i \neq j$$

Kinda...

Constructing conjugate directions $\mathbf{d}_j$: the home stretch

$\mathbf{d}_{j+1}^T \mathbf{H} \mathbf{d}_j = 0$ *(Part I)*

$\mathbf{d}_j^T \mathbf{H} \mathbf{d}_i = 0 \quad \Rightarrow \quad \mathbf{d}_{j+1}^T \mathbf{H} \mathbf{d}_i = 0, \ \forall i < j$ *(Part II)*

By induction:

$\mathbf{d}_2^T \mathbf{H} \mathbf{d}_1 = 0$ *(Part I)* $\quad$ $\mathbf{d}_3^T \mathbf{H} \mathbf{d}_1 = 0$ *(Part II)*

$\mathbf{d}_3^T \mathbf{H} \mathbf{d}_2 = 0$ *(Part I)* $\quad$ $\mathbf{d}_4^T \mathbf{H} \mathbf{d}_1 = 0$ *(Part II)*

$\mathbf{d}_4^T \mathbf{H} \mathbf{d}_2 = 0$ *(Part II)* $\quad$ $\mathbf{d}_4^T \mathbf{H} \mathbf{d}_3 = 0$ *(Part I)*

Etc., etc., etc...

So where are we now?

1. Choose an initial weight vector $\mathbf{w}_1$ and let $\mathbf{d}_1 = -\mathbf{g}_1$.
2. Update weight vector:

$$\mathbf{w}_{j+1} = \mathbf{w}_j + \alpha_j \mathbf{d}_j, \ j = \{1, 2, \ldots, W\}, \ \alpha_j = \frac{-\mathbf{d}_j^T \mathbf{g}_j}{\mathbf{d}_j^T \mathbf{H} \mathbf{d}_j}$$

3. Evaluate $\mathbf{g}_{j+1}$.
4. Let $\mathbf{d}_{j+1} = -\mathbf{g}_{j+1} + \beta_j \mathbf{d}_j$ where $\beta_j = \frac{\mathbf{g}_{j+1}^T \mathbf{H} \mathbf{d}_j}{\mathbf{d}_j^T \mathbf{H} \mathbf{d}_j}$
5. Let $j = j + 1$ and go to step 2.

What's the problem?

So where are we now?

Remaining Big Question:

- How can we do everything without computing $\mathbf{H}$?

Two areas:

$$\beta_j = \frac{\mathbf{g}_{j+1}^T \mathbf{H} \mathbf{d}_j}{\mathbf{d}_j^T \mathbf{H} \mathbf{d}_j}$$

$$\alpha_j = \frac{-\mathbf{d}_j^T \mathbf{g}_j}{\mathbf{d}_j^T \mathbf{H} \mathbf{d}_j}$$

Computing β_j without $\mathbf{H}$

$$\beta_j = \frac{\mathbf{g}_{j+1}^T \mathbf{H} \mathbf{d}_j}{\mathbf{d}_j^T \mathbf{H} \mathbf{d}_j}$$

From earlier:

$$\mathbf{H} \mathbf{d}_j = \frac{1}{\alpha_j} (\mathbf{g}_{j+1} - \mathbf{g}_j)$$

So:

$$\beta_j = \frac{\mathbf{g}_{j+1}^T (\mathbf{g}_{j+1} - \mathbf{g}_j)}{\mathbf{d}_j^T (\mathbf{g}_{j+1} - \mathbf{g}_j)} \quad \textit{(Hestenes-Stiefel)}$$

or...

Computing β_j without $\mathbf{H}$

$$\mathbf{d}_{j+1} = -\mathbf{g}_{j+1} + \beta_j \mathbf{d}_j$$

$$\mathbf{d}_j = -\mathbf{g}_j + \beta_{j-1}\mathbf{d}_{j-1}$$

Transpose and post-multiply by $\mathbf{g}_j$:

$$\mathbf{d}_j^T \mathbf{g}_j = -\mathbf{g}_j^T \mathbf{g}_j + \beta_{j-1}\mathbf{d}_{j-1}^T \mathbf{g}_j$$

Since:

$$\mathbf{d}_k^T \mathbf{g}_j = 0\,,\ \forall k < j$$

$$\mathbf{d}_j^T \mathbf{g}_j = -\mathbf{g}_j^T \mathbf{g}_j$$

$$\mathbf{g}_j^T \mathbf{g}_j = -\mathbf{d}_j^T \mathbf{g}_j$$

Computing β_j without $\mathbf{H}$

$$\beta_j = \frac{\mathbf{g}_{j+1}^T(\mathbf{g}_{j+1} - \mathbf{g}_j)}{\mathbf{d}_j^T(\mathbf{g}_{j+1} - \mathbf{g}_j)}$$

$$\mathbf{d}_k^T \mathbf{g}_j = 0\,,\ \forall k < j$$

$$\mathbf{g}_j^T \mathbf{g}_j = -\mathbf{d}_j^T \mathbf{g}_j$$

So:

$$\beta_j = \frac{\mathbf{g}_{j+1}^T(\mathbf{g}_{j+1} - \mathbf{g}_j)}{\mathbf{g}_j^T \mathbf{g}_j} \quad \textit{(Polak-Ribiere)}$$

Computing β_j without $\mathbf{H}$

$$\beta_j = \frac{\mathbf{g}_{j+1}^T \mathbf{g}_{j+1}}{\mathbf{g}_j^T \mathbf{g}_j}$$

$$\mathbf{g}_k^T \mathbf{g}_j = 0\,,\ \forall k < j$$

So:

$$\beta_j = \frac{\mathbf{g}_{j+1}^T \mathbf{g}_{j+1}}{\mathbf{g}_j^T \mathbf{g}_j} \quad \textit{(Fletcher-Reeves)}$$

Computing β_j without $\mathbf{H}$

Three choices:

$$\beta_j = \frac{\mathbf{g}_{j+1}^T(\mathbf{g}_{j+1} - \mathbf{g}_j)}{\mathbf{d}_j^T(\mathbf{g}_{j+1} - \mathbf{g}_j)} \quad \textit{(Hestenes-Stiefel)}$$

$$\beta_j = \frac{\mathbf{g}_{j+1}^T(\mathbf{g}_{j+1} - \mathbf{g}_j)}{\mathbf{g}_j^T \mathbf{g}_j} \quad \textit{(Polak-Ribiere)}$$

$$\beta_j = \frac{\mathbf{g}_{j+1}^T \mathbf{g}_{j+1}}{\mathbf{g}_j^T \mathbf{g}_j} \quad \textit{(Fletcher-Reeves)}$$

Which is best?

Computing α_j without $\mathbf{H}$

Key: Replace,

$$\alpha_j = \frac{-\mathbf{d}_j^T \mathbf{g}_j}{\mathbf{d}_j^T \mathbf{H} \mathbf{d}_j}$$

with line minimization.

Computing α_j without $\mathbf{H}$

$$E(\mathbf{w}) = E_0 + \mathbf{b}^T \mathbf{w} + \frac{1}{2} \mathbf{w}^T \mathbf{H} \mathbf{w}$$

So:

$$E(\mathbf{w}_j + \alpha_j \mathbf{d}_j) = E_0 + \mathbf{b}^T (\mathbf{w}_j + \alpha_j \mathbf{d}_j) + \frac{1}{2} (\mathbf{w}_j + \alpha_j \mathbf{d}_j)^T \mathbf{H} (\mathbf{w}_j + \alpha_j \mathbf{d}_j)$$

$$\partial E(\mathbf{w}_j + \alpha_j \mathbf{d}_j) / \partial \alpha_j = \mathbf{b}^T \mathbf{d}_j + (1/2) \mathbf{d}_j^T \mathbf{H} (\mathbf{w}_j + \alpha_j \mathbf{d}_j) + (1/2) (\mathbf{w}_j + \alpha_j \mathbf{d}_j)^T \mathbf{H} \mathbf{d}_j$$

Computing α_j without $\mathbf{H}$

$$\partial E(\mathbf{w}_j + \alpha_j \mathbf{d}_j) / \partial \alpha_j = \mathbf{b}^T \mathbf{d}_j + (1/2) \mathbf{d}_j^T \mathbf{H} (\mathbf{w}_j + \alpha_j \mathbf{d}_j) + (1/2) (\mathbf{w}_j + \alpha_j \mathbf{d}_j)^T \mathbf{H} \mathbf{d}_j$$

Since $\mathbf{H}$ is symmetric:

$$\mathbf{d}_j^T \mathbf{H} (\mathbf{w}_j + \alpha_j \mathbf{d}_j) = (\mathbf{w}_j + \alpha_j \mathbf{d}_j)^T \mathbf{H} \mathbf{d}_j$$

$$\partial E(\mathbf{w}_j + \alpha_j \mathbf{d}_j) / \partial \alpha_j = \mathbf{b}^T \mathbf{d}_j + \mathbf{d}_j^T \mathbf{H} (\mathbf{w}_j + \alpha_j \mathbf{d}_j) = 0$$

$$\mathbf{b}^T \mathbf{d}_j + \mathbf{d}_j^T \mathbf{H} (\mathbf{w}_j + \alpha_j \mathbf{d}_j) = 0$$

$$\mathbf{d}_j^T \mathbf{b} + \mathbf{d}_j^T \mathbf{H} \mathbf{w}_j + \alpha_j \mathbf{d}_j^T \mathbf{H} \mathbf{d}_j = 0$$

Computing α_j without $\mathbf{H}$

$$\mathbf{d}_j^T \mathbf{b} + \mathbf{d}_j^T \mathbf{H} \mathbf{w}_j + \alpha_j \mathbf{d}_j^T \mathbf{H} \mathbf{d}_j = 0$$

So:

$$\alpha_j \mathbf{d}_j^T \mathbf{H} \mathbf{d}_j = -\mathbf{d}_j^T (\mathbf{b} + \mathbf{H} \mathbf{w}_j)$$

$$\mathbf{g}_j = \mathbf{b} + \mathbf{H} \mathbf{w}_j$$

$$\alpha_j \mathbf{d}_j^T \mathbf{H} \mathbf{d}_j = -\mathbf{d}_j^T \mathbf{g}_j$$

$$\alpha_j = \frac{-\mathbf{d}_j^T \mathbf{g}_j}{\mathbf{d}_j^T \mathbf{H} \mathbf{d}_j}$$

Conclusion: line minimization = α_j computation...

Complete conjugate gradient algorithm

1. Choose an initial weight vector $\mathbf{w}_1$ and let $\mathbf{d}_1 = -\mathbf{g}_1$.
2. Perform a line minimization along $\mathbf{d}_j$, such that:

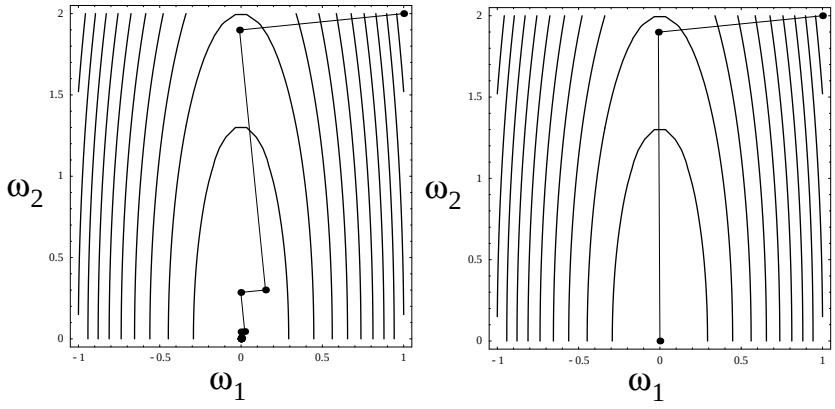
   $E(\mathbf{w}_j + \alpha^*\mathbf{d}_j) \leq E(\mathbf{w}_j + \alpha\mathbf{d}_j)$, $\forall\alpha$.
3. Let $\mathbf{w}_{j+1} = \mathbf{w}_j + \alpha^*\mathbf{d}_j$.
4. Evaluate $\mathbf{g}_{j+1}$.
5. Let $\mathbf{d}_{j+1} = -\mathbf{g}_{j+1} + \beta_j\mathbf{d}_j$ where,

   $$\beta_j = \frac{\mathbf{g}_{j+1}^T(\mathbf{g}_{j+1} - \mathbf{g}_j)}{\mathbf{g}_j^T\mathbf{g}_j}$$ *(Polak-Ribiere)*
6. Let $j = j + 1$ and go to step 2.

Comments

- **Exploitation of reasonable assumption about local quadratic nature of error surface.**
- **Little additional computation beyond steepest descent.**
- **No Hessian computation required.**
- **No hand-tuning of learning rate.**
- **In practice, conjugate gradient algorithm must be reset every W steps.** ***(Why?)***
- **What about violations of $\mathbf{H} > 0$ assumption?**

Quadratic example

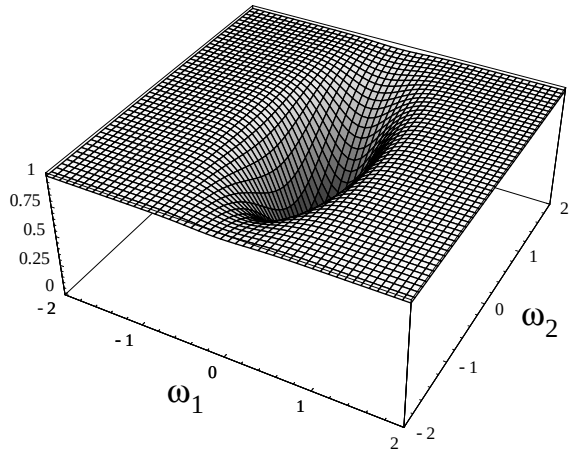


steepest descent
15 steps to convergence

conjugate gradient
2 steps to convergence

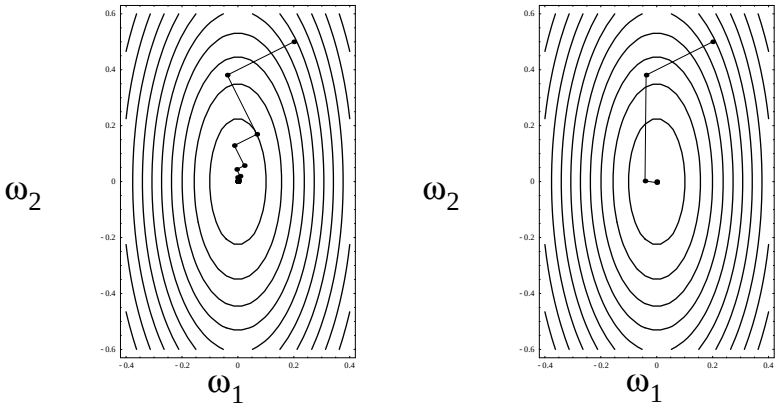
$E = 20\omega_1^2 + \omega_2^2$

Nonquadratic example



$E(\omega_1, \omega_2) = 1 - \exp(-5\omega_1^2 - \omega_2^2)$

Nonquadratic example



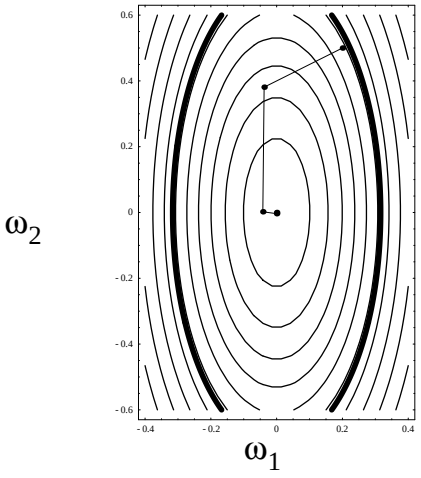
steepest descent
25 steps to convergence

conjugate gradient
4 steps to convergence

$(\omega_1, \omega_2) = (0.2, 0.5)$

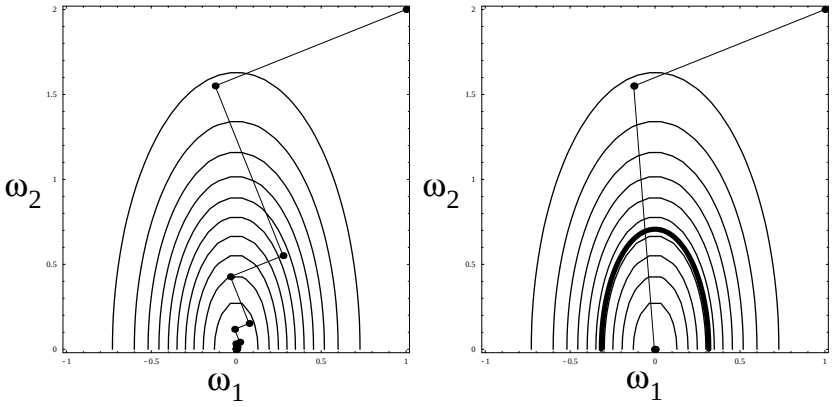
(why these initial weights?)

Nonquadratic example



Region where $\mathbf{H} > 0$

Nonquadratic example ($\mathbf{H} < 0$)



steepest descent
24 steps to convergence

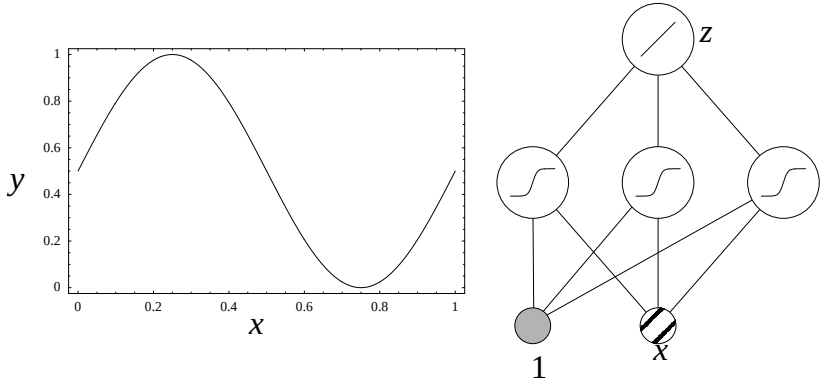
conjugate gradient
5 steps to convergence

$(\omega_1, \omega_2) = (1, 2)$

(gradient descent — 940 steps!)

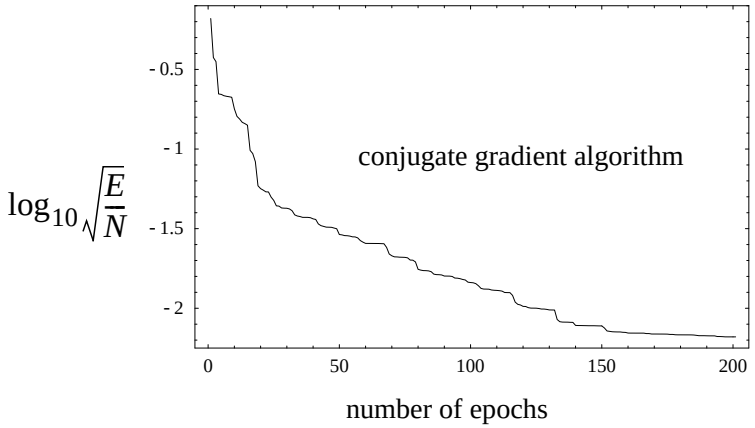
Simple NN training example

$$y = \frac{1}{2} + \frac{1}{2}\sin(2\pi x), \; 0 \le x \le 1$$



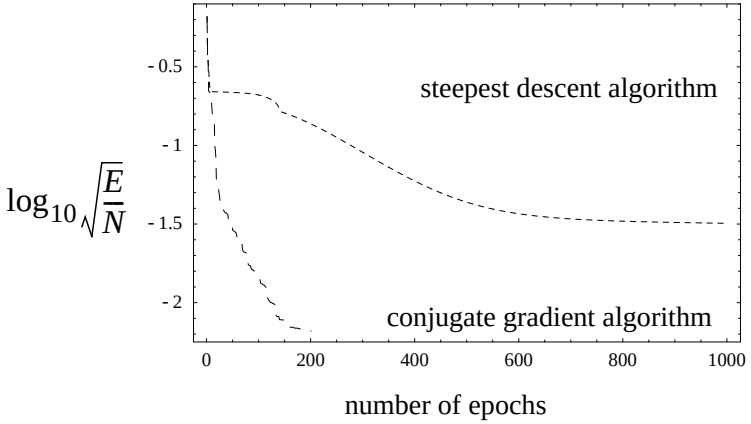
Simple NN training example

Error convergence



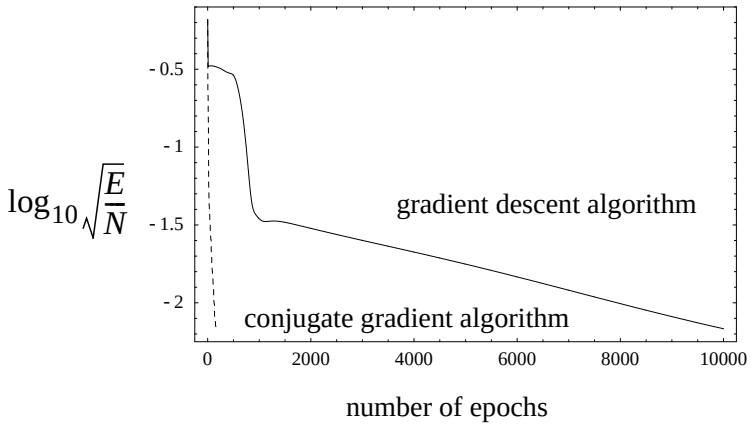
Simple NN training example

Error convergence

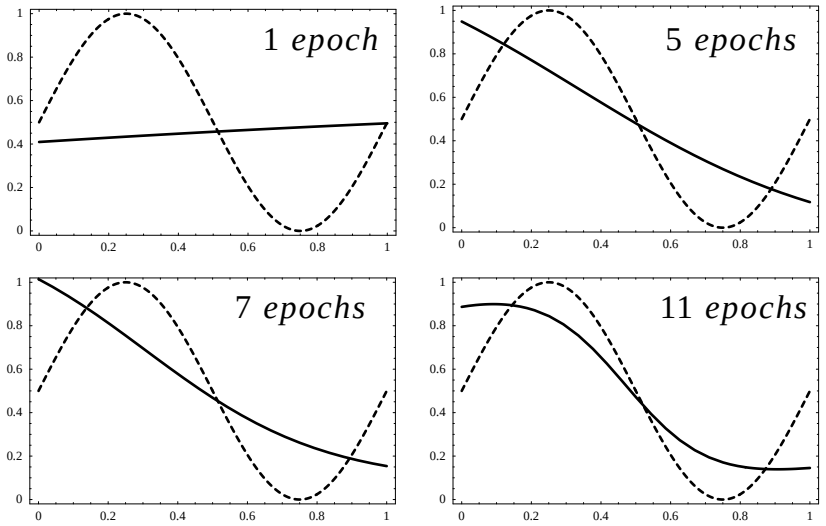


Simple NN training example

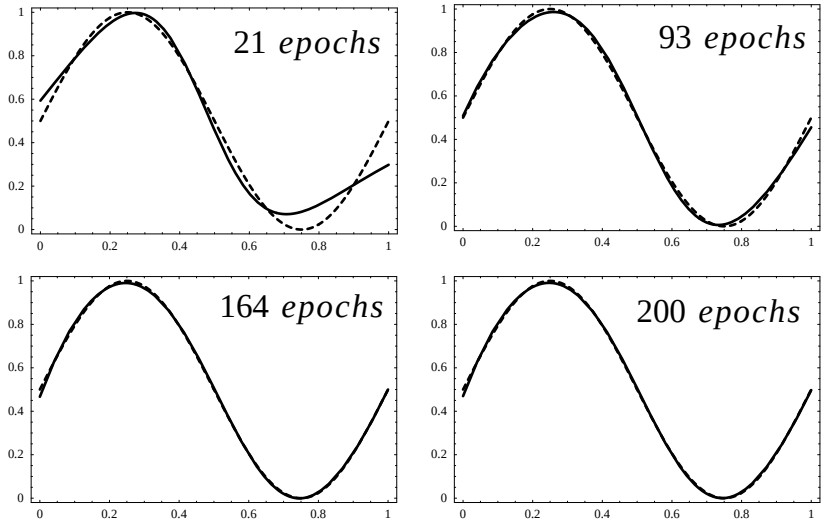
Error convergence



A closer look at convergence

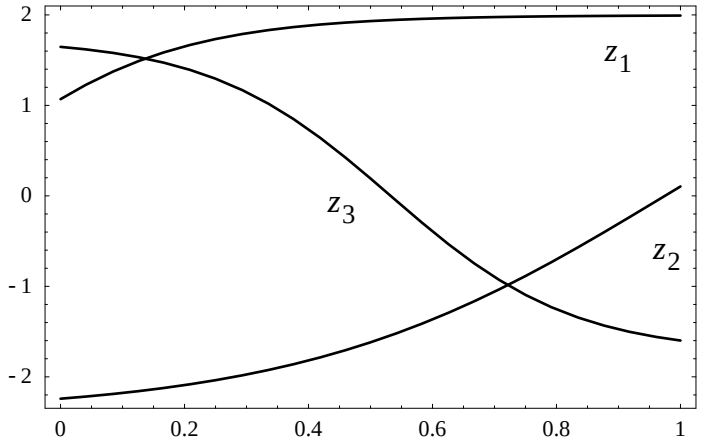


A closer look at convergence



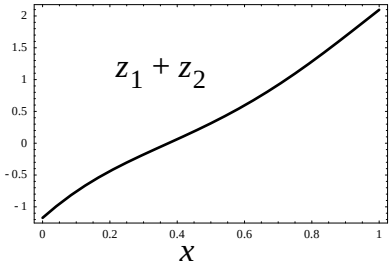
Final NN approximation: a closer look

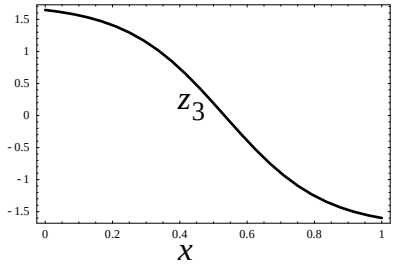
Hidden unit outputs (z_1, z_2 and z_3)



Final NN approximation: a closer look

Hidden unit outputs (z_1, z_2 and z_3)





Conjugate gradient conclusions

- **Exploitation of reasonable assumption about local quadratic nature of error surface.**
- **Little additional computation beyond steepest descent.**
- **No Hessian computation required.**

- ***No hand-tuning of learning rate.***
- ***Much faster rate of convergence.***