Conjugate Directions.

We've seen that neither coordinate directions* nor gradient descent are very effective when approaching a very "elliptical" minimum.

What's the problem? Observe

Lemma. If we minimize f along direction $\vec{d}$ from $\vec{x}$ at the minimum, $\nabla f \perp \vec{d}$.

Proof. $\langle \nabla f, \vec{d} \rangle$ is the derivative of f along the line. $\square$

Now we note that around any P,

$$f(\vec{p}+\vec{x}) = f(\vec{p}) + G(\vec{p}) \cdot \vec{x} + \frac{1}{2} \vec{x} H(\vec{p}) \vec{x} + \ldots$$

* Note that if your function is a sum or (more subtly) a product of functions of the individual variables $x_1, \dots, x_n$, then coordinate search works beautifully, solving the problem exactly in n steps.

So if $G(\vec{p}) = -b$, and $H(\vec{p}) = A$, (notice me!)
we can compute

$$\nabla f(\vec{p}+\vec{x}) = -b + A\vec{x}.$$

Now we observe that this implies we can find the $\vec{x}$ so that $\nabla f(\vec{p}+\vec{x})=0$ by solving

$$A\vec{x} - \vec{b} = 0.$$

(and there are methods which do this!)

Now how does ∇f change as we move in direction $\vec{v}$? Well, we expect

$$\Delta(\nabla f) \approx \Delta(-b + A\vec{x})$$
$$= A\vec{v}$$

Now when we minimized along $\vec{u}$, we arrived at a point where $\langle \nabla f, \vec{u} \rangle = 0$. In our search for a point where $\nabla f = 0$, we'd like to keep the property that $\langle \nabla f, \vec{u} \rangle = 0$ even when we move along $\vec{v}$. For that to work, we need the change

$$D_{\vec{v}} \left(\langle \nabla f, \vec{u} \rangle \right) = 0$$

or

$$\langle \vec{u}, D_{\vec{v}} (\nabla f) \rangle = 0$$

or (approximately)

$$\langle \vec{u}, A \vec{v} \rangle = 0$$

Definition. We say that $\vec{u}$ and $\vec{v}$ are conjugate (with respect to A) if $\langle \vec{u}, A\vec{v} \rangle = 0$.

Note: If A is positive-definite (as it is at a nondegenerate minimum of f!), we note that this is just the condition that $\vec{u}$ and $\vec{v}$ are orthogonal <u>in the metric generated by A</u>.

Meta-algorithm. Generate a set of n (linearly independent) conjugate directions and minimize along each in turn.

Fletcher, Theorem 2.4.1.

Proposition. If f is quadratic and the Hessian of f is positive definite, the meta-algorithm finds the exact min (in n steps).

Proof. In general, $\exists$ some $\vec{x}'$ and c' so that any quadratic function of $\vec{x}$ can be written

$$q(\vec{x}) = \frac{1}{2}(\vec{x}-\vec{x}')^T A(\vec{x}-\vec{x}') + c'.$$

It's not hard to see that

1) A is the Hessian of q

2) $\vec{x}'$ is the minimizer of q if A is pos. def.

Suppose $\vec{s}_1, \ldots, \vec{s}_n$ are a set of conjugate directions. We first show that the $\vec{s}_i$ form a basis for $\mathbb{R}^n$.

Suppose (wlog) not. Then

$$\vec{s}_1 = \sum_{i=2}^{n} a_i \vec{s}_i,$$

where $a_2 \neq 0$. (we could always reorder to get this).

$$\vec{s}_1^{\,T} A \vec{s}_2 = \left(\sum a_i \vec{s}_i\right)^T A \vec{s}_2$$

$$= \underbrace{(a_2 \vec{s}_2)^T A \vec{s}_2}_{\substack{\neq 0, \text{ since} \\ a_2 \neq 0, \ A \text{ is} \\ \text{pos. def.}}} + \underbrace{\sum_{i=3}^{n} (a_i \vec{s}_i)^T A \vec{s}_2}_{\substack{0, \text{ since the} \\ \vec{s}_i \text{ are conjugate} \\ \text{to } \vec{s}_2}}$$

On the other hand,

$$\vec{s}_1^{\,T} A \vec{s}_2 = 0, \quad \text{since } s_1 \text{ is conjugate to } s_2.$$

This means that if we start the meta-algorithm at $\vec{x}^{(1)}$ then we can write the mimimizer as

$$\vec{x}' = \vec{x}^{(1)} + \sum a_i' \vec{s}_i$$

and any other point

$$\vec{x} = \vec{x}^{(1)} + \sum' a_i \vec{s}_i .$$

Thus we can write q in terms of the vector $\vec{a} = (a_1, \ldots, a_n)$ as

$$q(\vec{a}) = \frac{1}{2}(\vec{a} - \vec{a}')S^T A S(\vec{a} - \vec{a}') + c'.$$

where $S = \begin{pmatrix} \uparrow & \uparrow & & \uparrow \\ \vec{s}_1 & \vec{s}_2 & \cdots & \vec{s}_n \\ \downarrow & \downarrow & & \downarrow \end{pmatrix}$. Now the $\vec{s}_i$ are conjugate, so

$$S^T A S = D \leftarrow \text{a diagonal matrix with diagonal entries } d_i = \vec{s}_i^{\,T} A \vec{s}_i .$$

so

$$q(\vec{a}) = \frac{1}{2} \sum' (a_i - a_i')^2 d_i$$

Now line search along the ~~e~~ $\vec{s}_i$ directions amounts to coordinate search on the a_i coordinates, which is guaranteed to converge, since $q(\vec{a})$ is the sum of functions of the individual a_i. $\square$

Clever! But how can we take advantage of this method? If we knew the Hessian, we could generate conjugate directions by Gram-Schmidt relative to the metric generated by the Hessian. But usually, we don't know the Hessian, so we need a little more cleverness.

This will lead us directly to the "conjugate-gradient" methods, which are based on the following theorem.

Orthogonalization Theorem.

Suppose A is a symmetric positive definite matrix. Let $\vec{g}_0$ be any vector and $\vec{h}_0 = \vec{g}_0$. Define a sequence of vectors $\vec{h}_i, \vec{g}_i$ by

$$\vec{g}_{i+1} = \vec{g}_i - \lambda_i A \vec{h}_i$$
$$\vec{h}_{i+1} = \vec{g}_{i+1} - \gamma_i \vec{h}$$

where we choose λ_i and γ_i so that $\vec{g}_{i+1} \cdot \vec{g}_i = 0$, $\vec{h}_{i+1}^T A \vec{h}_i = 0$. Then for all $i \leq n$, $\{g_0, \ldots, g_i\}$ is a set of <u>mutually</u> orthogonal vectors and

$\{h_0, \dots, h_i\}$ is a set of <u>mutally</u> conjugate vectors.

We note that it's easy to see

$$\lambda_i = \frac{\vec{g}_i \cdot \vec{g}_i}{\vec{g}_i^T A \vec{h}_i}, \quad \gamma_i = \frac{\vec{g}_{i+1}^T A \vec{h}_i}{\vec{h}_i^T A \vec{h}_i}.$$

Check:

$$\vec{g}_{i+1} \cdot \vec{g}_i = \left(\vec{g}_i - \frac{\vec{g}_i \cdot \vec{g}_i}{\vec{g}_i^T A \vec{h}_i} A \vec{h}_i \right) \cdot \vec{g}_i$$

$$= \vec{g}_i \cdot \vec{g}_i - \vec{g}_i \cdot \vec{g}_i = 0.$$

$$\vec{h}_{i+1}^T A \vec{h}_i = \left(\vec{g}_{i+1} + \left(\frac{\vec{g}_{i+1}^T A \vec{h}_i}{\vec{h}_i^T A \vec{h}_i} \right) \vec{h}_i \right)^T A \vec{h}_i$$

$$= \vec{g}_{i+1}^T A \vec{h}_i - \vec{g}_{i+1}^T A \vec{h}_i$$

$$= 0.$$

Corollary. We claim

$$\gamma_i = -\frac{\vec{g}_{i+1}\cdot\vec{g}_{i+1}}{\vec{g}_i\cdot\vec{g}_i} = -\frac{(\vec{g}_{i+1}-\vec{g}_i)\cdot\vec{g}_{i+1}}{\vec{g}_i\cdot\vec{g}_i}$$

and

$$\lambda_i = \frac{\vec{g}_i\cdot\vec{h}_i}{\vec{h}_i^T A\vec{h}_i}.$$

Proof. Recall that

$$\vec{g}_{i+1} = \vec{g}_i - \lambda_i A\vec{h}_i$$

so $\dfrac{\vec{g}_{i+1}-\vec{g}_i}{-\lambda_i} = A\vec{h}_i$. Thus

$$\gamma_i = \frac{\vec{g}_{i+1}(A\vec{h}_i)}{\vec{h}_i^T A\vec{h}_i} = \frac{\vec{g}_{i+1}\cdot(\vec{g}_{i+1}-\vec{g}_i)}{-\lambda_i\vec{h}_i^T A\vec{h}_i}$$

Now

$$-\lambda_i\vec{h}_i^T A\vec{h}_i = \frac{-\vec{g}_i\cdot\vec{g}_i}{\vec{g}_i^T A\vec{h}_i}\vec{h}_i^T A\vec{h}_i$$

We now consider

$$\frac{h_i^T A h_i}{g_i^T A h_i}$$

Now

$$\vec{g}_i = \vec{h}_i - \gamma_{i-1} \vec{h}_{i-1}$$

so by conjugacy of the h_i, we have

$$\vec{g}_i^T A \vec{h}_i = \vec{h}_i^T A \vec{h}_i .$$

Thus we get

$$\gamma_i = \frac{\vec{g}_{i+1} \cdot (\vec{g}_{i+1} - \vec{g}_i)}{- \vec{g}_i \cdot \vec{g}_i} = - \frac{\vec{g}_{i+1} \cdot \vec{g}_{i+1}}{\vec{g}_i \cdot \vec{g}_i} .$$

We now tackle

$$\lambda_i = \frac{\vec{g}_i \cdot \vec{g}_i}{\vec{g}_i^T A \vec{h}_i} = \cancel{\frac{\vec{g}_i \cdot \vec{g}_i (1 - \gamma_i)}{\vec{g}_i (\vec{g}_{i+1} - \vec{g}_i)}}$$

Now we see

$$\vec{g}_i^T(A\vec{h}_i) = (\vec{h}_i - \gamma_{i-1}\vec{h}_{i-1})^T A \vec{h}_i$$
$$= \vec{h}_i^T A \vec{h}_i$$

and we have

$$\vec{g}_i \cdot \vec{g}_i = \vec{g}_i \cdot (\vec{h}_i - \gamma_{i-1}\vec{h}_{i-1})$$
$$= \vec{g}_i \cdot \vec{h}_i - \gamma_{i-1}(\vec{g}_i \cdot \vec{h}_{i-1}).$$

But $\vec{h}_{i-1} = \vec{g}_{i-1} - \gamma_{i-2}\vec{h}_{i-2}$ so we can get rid of the last term by descent (recall $h_0 = g_0$, which is orthogonal to all the other g_i).

Thus

$$\lambda_i = \frac{\vec{g}_i \cdot \vec{h}_i}{\vec{h}_i^T A \vec{h}_i},$$ as claimed. $\square$

So what? We can construct γ_i without A, but our formula for λ_i still involves A.

Proposition. Suppose $\vec{g}_i = -\nabla f(\vec{P}_i)$ and we proceed in direction $\vec{h}_i$ to the local min at $\vec{P}_{i+1}$, and let $g_{i+1} = -\nabla f(P_{i+1})$. Then $g_{i+1} = g_i - \lambda_i A h_i$, following our procedure above.

Proof. We saw last class that if f is quadratic, then $f(\vec{x}) = c - \vec{b}\cdot\vec{x} + \frac{1}{2}\vec{x}^T A\vec{x}$,

$$\nabla f(\vec{x}) = A\vec{x} - \vec{b}.$$

We then have (where A is the Hessian)

$$g_i = -\nabla f(P_i) = -A\cdot P_i + b$$

and

$$(*) \quad g_{i+1} = -\nabla f(P_{i+1})$$
$$= -\nabla f(P_i + \lambda \vec{h}_i)$$
$$= -A(P_i + \lambda \vec{h}_i) + b$$
$$= \vec{g}_i - \lambda A \vec{h}_i .$$

where we chose λ to minimize f along the line in direction $\vec{h}_i$ from $\vec{P}_i$. But that is exactly the λ which makes

$$g_{i+1} = -\nabla f(P_{i+1}) \perp \text{ to } h_i,$$

or

$$g_{i+1} \cdot h_i = 0.$$

or

$$0 = \vec{h}_i \cdot (\vec{g}_i - \lambda A \vec{h}_i)$$
$$= \vec{h}_i \cdot \vec{g}_i - \lambda \vec{h}_i A \vec{h}_i$$

or (wonderfully!)

$$\lambda = \frac{\vec{h}_i \cdot \vec{g}_i}{\vec{h}_i^T A \vec{h}_i} = \lambda_i . \qquad \square$$

Thus, just by minimizing in direction h_i and reading off the g_i from ∇f, we can reconstruct the entire sequence of g_i and h_i without ever explictly Knowing A!

We now can outline the conjugate gradient algorithm.

Polak-Ribiere Conjugate Gradient Algorithm.

Set $p = x_0$ $g = -\nabla f(p)$, $h = \xi = g$.

main loop:

minimize from p in direction ξ.

if reduction is small enough, return

$f_p = f(p)$, $\xi = \nabla f(p)$.

$$\gamma = \frac{(\xi + g)\cdot \xi}{g \cdot g} \quad \left(\gamma = \frac{-(g_{i+1} - g_i)\cdot g_{i+1}}{g_i \cdot g_i}\right)$$

$g = -\xi$

$\xi = h = g + \gamma h$

go back to top of loop.

We note that we could also have used $\gamma = -\dfrac{\vec{g}_{i+1}\cdot \vec{g}_{i+1}}{\vec{g}_i \cdot \vec{g}_i}$ (from our corollary)

For a function that's <u>exactly</u> quadratic, this would make no difference. But for a non-quadratic function, if we need to do another round of iterations, the Polak-Ribiere version is reported to work better in practice (the other version is called Fletcher-Reeves).

<mathematica demo>