More gradient desent

Let's try another, more important example to illustrate the behavior of gradient desent. Recall:

$$x_{n+1} = x_n - \alpha_n \nabla f(x_n), \qquad \alpha_n > 0.$$

Stick w/ $\alpha_n \equiv 1$ for now.

Let's try the ~~dest~~ cost function:

$$f(x,y) = ax^2 + by^2.$$

Then:

$$\nabla f(x,y) = (2ax, 2by).$$

Clearly, the optimum is:

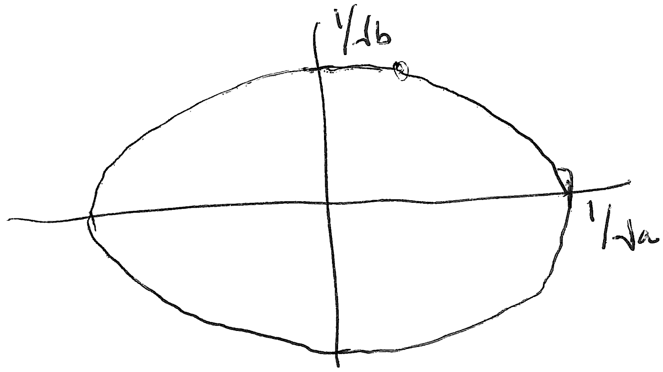
$$(x^*, y^*) = (0,0).$$

Note that this is a quadratic form with a particularly simple set of coefficients:

$$f(p) = p^T A p, \quad p = \begin{bmatrix} x \\ y \end{bmatrix}, \quad A = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}.$$

What does the 1-level set of f look like? Well, $1 = ax^2 + by^2$ is an ellipse with major and minor radii $\frac{1}{\sqrt{a}}$ and $\frac{1}{\sqrt{b}}$.

Looks like;



Let's consider the line starting at (x_0, y_0), with direction given by $\nabla f(x_0, y_0) = (2ax_0, 2by_0)$. The symmetric equations for this line are:

$$2ax_0(x - x_0) = 2by_0(y - y_0).$$

Note that for general (x_0, y_0), there is no choice of (a, b) which satisfies these equations! What is happening?

Let's see what happens if we apply the exact line search to this problem. Set:

$$g(\alpha) = f(x_k - 2\alpha a x_k, y_k - 2\alpha b y_k)$$

and solve:

$$\underset{\alpha > 0}{\text{minimize}} \; g(\alpha).$$

We have:

$$g'(\alpha) = \frac{d}{d\alpha}\left\{ a(x_k - 2\alpha a x_k)^2 + b(y_k - 2\beta b y_k)^2 \right\}$$

$$= -2a^2 x_k (x_k - 2\alpha a x_k) - 2b^2 y_k (y_k - 2\alpha b y_k)$$

$$\Rightarrow \quad 0 = a^2 x_k^2 (1 - 2\alpha a) + b^2 y_k^2 (1 - 2\alpha b)$$

$$\Rightarrow \quad \alpha_k = \frac{a^2 x_k^2 + b^2 y_k^2}{2(a^3 x_k^2 + b^3 y_k^2)}$$

So, we can compute α_k exactly in this case. Makes sense, cost function is quadratic, search path is linear, composition of quadratic & linear is quadratic, & derivative is linear. Should be able to solve.

We can expet if we iterate:

$$x_{k+1} = x_k - \alpha_k f_x(x_k, y_k) = x_k - 2\alpha_k a x_k$$

$$y_{k+1} = \quad \ldots \quad = y_k - 2\alpha_k b y_k.$$

What does this look like?

$\boxed{\text{Show demo.}}$

What if we try to solve this using Newton's method?

Need to compute the Hessian. Recall:

$$\nabla^2 f(x,y) = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x^2} & \dfrac{\partial^2 f}{\partial x \partial y} \\[2ex] \dfrac{\partial^2 f}{\partial y \partial x} & \dfrac{\partial^2 f}{\partial^2 y} \end{bmatrix}_{(x,y)}$$

we have:

$$\frac{\partial^2 f}{\partial x^2} = 2a, \qquad \frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x} = 0, \qquad \frac{\partial^2 f}{\partial y^2} = 2b.$$

Hence, Newton's method requires us to apply:

$$\nabla^2 f(x_k, y_k)^{-1} \equiv \begin{bmatrix} 2a & 0 \\ 0 & 2b \end{bmatrix}^{-1} = \frac{1}{2}\begin{bmatrix} a^{-1} & 0 \\ 0 & b^{-1} \end{bmatrix}.$$

inverse of a diagonal matrix is matrix with reciprocal of diagonal elements

Hence:

$$\underline{p}_{k+1} = \underline{p}_k - \nabla^2 f(x_k, y_k) \nabla f(x_k, y_k)$$

$$= \underline{p}_k - \frac{1}{2}\begin{bmatrix} a^{-1} & 0 \\ 0 & b^{-1} \end{bmatrix}\begin{bmatrix} 2a x_k \\ 2b y_k \end{bmatrix} = p_k - \frac{1}{2}\begin{bmatrix} 2a^{-1} a x_k \\ 2b^{-1} b y_k \end{bmatrix}$$

$$= p_k - \begin{bmatrix} x_k \\ y_k \end{bmatrix} = p_k - p_k = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} x^* \\ y^* \end{bmatrix}!$$

From any starting point, we minimize exactly in one step using Newton's method!

why is this happening?

Let's consider a more general quadratic form:

$$q(\underline{x}) = \frac{1}{2}\underline{x}^T A \underline{x} + \underline{b}^T \underline{x} + c.$$

$$= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij} x_i x_j + \sum_{i=1}^{n} b_i x_i + c.$$

Let's compute some partial derivatives:

$$\frac{\partial q}{\partial x_k} = \underbrace{\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}\frac{\partial}{\partial x_k} x_i x_j}_{(*)} + \underbrace{\sum_{i=1}^{n} b_i \frac{\partial}{\partial x_k}}_{(**)} + 0$$

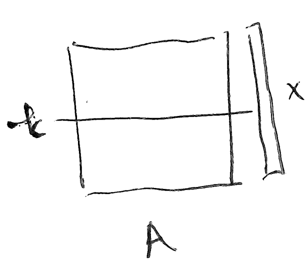
Let's do each of these terms carefully:

$$(*): \quad \frac{1}{2}\sum_{i=1}^{n}\frac{\partial}{\partial x_k}\left\{x_i \sum_{j=1}^{n} A_{ij} x_j\right\}$$

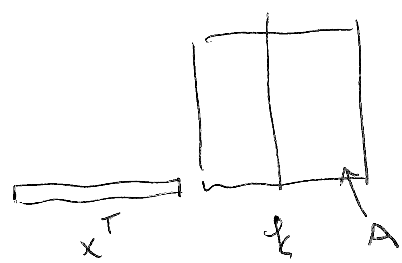
$$= \frac{1}{2}\sum_{i=1}^{n}\left\{\delta_{ik}\cdot\sum_{j=1}^{n} A_{ij} x_j + x_i \sum_{j=1}^{n} A_{ij}\delta_{jk}\right\}$$

$$= \frac{1}{2}\sum_{i=1}^{n}\delta_{ik}\sum_{j=1}^{n} A_{ij} x_j + \frac{1}{2}\sum_{i=1}^{n} x_i \sum_{j=1}^{n} A_{ij}\delta_{jk}$$

$$= \underbrace{\frac{1}{2}\sum_{j=1}^{n} A_{kj} x_j}_{} + \underbrace{\frac{1}{2}\sum_{i=1}^{n} x_i A_{ik}}_{}$$



A



$$= \frac{1}{2}\underbrace{A_{k,:}}_{(k^{th}\text{ row})} x + \frac{1}{2}\underbrace{(A_{:,k})^T}_{(k^{th}\text{ column})} x \quad \Leftarrow \text{"MATLAB notation"}$$

Easier term:

$$(**) \qquad \frac{\partial}{\partial x_k} \sum_{i=1}^{n} b_i x_i = \sum_{i=1}^{n} b_i \frac{\partial x_i}{\partial x_k} = \sum_{i=1}^{n} b_i \delta_{ik} = b_k .$$

What can we conclude?

$$\nabla q(x) = \frac{1}{2}(A + A^T)x + b.$$

Note: if $A = A^T$, $\nabla q(x) = Ax + b$.

What about the Hessian? Should be easy to see now that:

$$\nabla^2 q(x) = D\left\{ \frac{1}{2}(A + A^T)x + b \right\} = \frac{1}{2}(A + A^T).$$

Recall, for optimality, need:

$$\nabla q(x^*) = 0 \iff \frac{1}{2}(A + A^T)x^* + b = 0$$

$$\iff x^* = -2(A + A^T)^{-1} b.$$

For the Newton iteration:

$$\begin{aligned}
x_{k+1} &= x_k - \nabla^2 q(x_k)^{-1} \nabla q(x_k) \\
&= x_k - \left[\frac{1}{2}(A + A^T)\right]^{-1}\left(\frac{1}{2}(A + A^T)x_k + b\right) \\
&= x_k - x_k - 2(A + A^T)^{-1} b \\
&= -2(A + A^T)^{-1} b = x^*.
\end{aligned}$$

Whether it's a minimizer, or a maximizer, or whether it's unique depends on A...

So, again, we find a <u>stationary point</u> in one step...