Optimization

Class 8: Quasi-Newton Methods

Problems with Newton's optimization method:

- finds stationary points and not necessarily the optimizer you're looking for.
- can fail to work (e.g., can't invert second-derivative matrix).
- can be relatively expensive (in time at each step and storage) to carry out.

1 Favoring optimizers

Rather than testing the second-derivative matrix for positive-definiteness afterward, replace it in the update with a matrix that is certainly positive-definite:

$$\begin{bmatrix} x_{\text{new}} \\ y_{\text{new}} \end{bmatrix} = \begin{bmatrix} x_{\text{old}} \\ y_{\text{old}} \end{bmatrix} - \underbrace{\cancel{\nabla^2 f(x_{\text{old}}, y_{\text{old}})}}_{P}{}^{-1} \nabla f(x_{\text{old}}, y_{\text{old}})$$

1.1 Steepest descent

The simplest positive-definite matrix is the identity

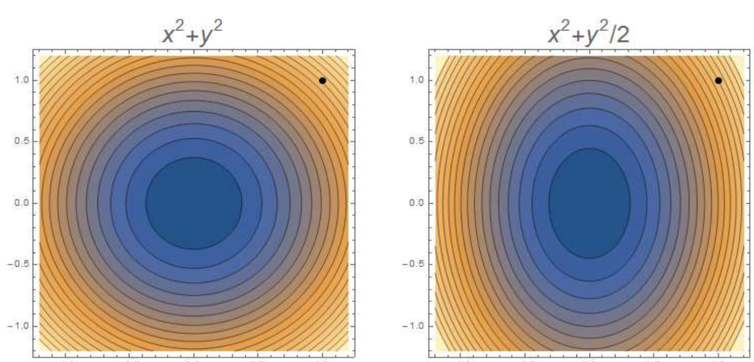
$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

This leads to the *steepest-descent* update formula

$$\begin{bmatrix} x_{\text{new}} \\ y_{\text{new}} \end{bmatrix} = \begin{bmatrix} x_{\text{old}} \\ y_{\text{old}} \end{bmatrix} - \nabla f(x_{\text{old}}, y_{\text{old}}),$$

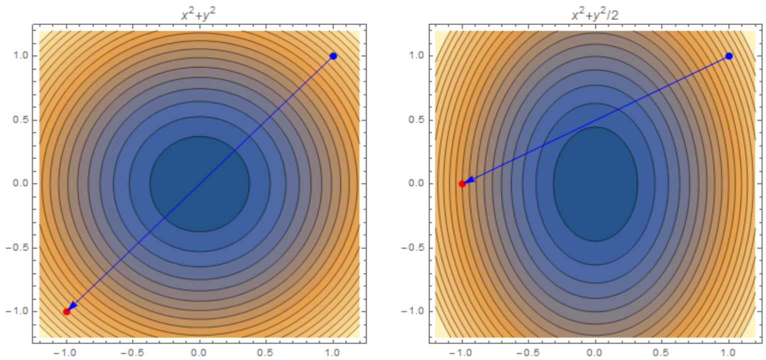
so-called because $-\nabla f(x_{\text{old}}, y_{\text{old}})$ points in the direction of steepest descent from the old guess.

Group Problem 1



Track one step of the steepest descent method applied to each function from the current guess $(x_{\text{old}}, y_{\text{old}}) = (1, 1)$.

Group Problem 1 (solution)



The gradient for the function pictured on the left is $2\,x\,\vec{i} + 2\,y\,\vec{j}$, which evaluates at $(x_{\text{old}}, y_{\text{old}}) = (1, 1)$ to be $2\,\vec{i} + 2\,\vec{j}$. Following the negative gradient from $(x_{\text{old}}, y_{\text{old}}) = (1, 1)$ then yields $(x_{\text{new}}, y_{\text{new}}) = (-1, -1)$.

The gradient for the function pictured on the right is $2\,x\,\vec{i} + y\,\vec{j}$, which evaluates at $(x_{\text{old}}, y_{\text{old}}) = (1, 1)$ to be $2\,\vec{i} + \vec{j}$. Following the negative gradient from $(x_{\text{old}}, y_{\text{old}}) = (1, 1)$ then yields $(x_{\text{new}}, y_{\text{new}}) = (-1, 0)$.

Both of these cases show that the steepest-descent method can overshoot a lower f-valued location along the step-direction, and the first case shows that following the steepest-descent direction too far might not even lead to descent (it can even lead to *ascent* eventually).

See Steepest Descent in Class8.nb.

2 Line-Searches

Instead, we can use the update

$$\begin{bmatrix} x_{\text{new}} \\ y_{\text{new}} \end{bmatrix} = \begin{bmatrix} x_{\text{old}} \\ y_{\text{old}} \end{bmatrix} - \alpha\, P^{-1}\, \nabla f(x_{\text{old}}, y_{\text{old}})$$

for the *step-factor* $\alpha \geq 0$, which allows us to ensure that we have a lower f-value at the new guess.

2.1 Backtracking line-search

This simple scheme starts at each step with a step-factor of $\alpha = 1$, and halves this repeatedly until actual descent in that direction is achieved.

2.2 Exact line-search

This scheme finds the step-factor that minimizes the objective function f along the step-direction at each step. For example, if we use the steepest descent step-direction:

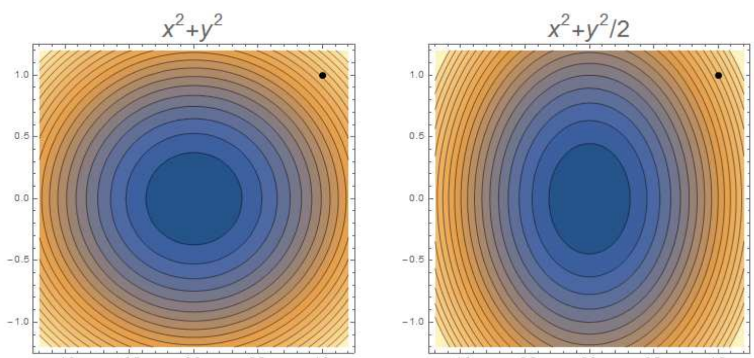
steepest descent with exact line-search

$$\begin{bmatrix} x_{\text{old}} \\ y_{\text{old}} \end{bmatrix} - \alpha \nabla f(x_{\text{old}}, y_{\text{old}}) = \begin{bmatrix} x_{\text{old}} - \alpha\, f_x(x_{\text{old}}, y_{\text{old}}) \\ y_{\text{old}} - \alpha\, f_y(x_{\text{old}}, y_{\text{old}}) \end{bmatrix}$$

$$\Downarrow$$

$$\min_{\alpha} \text{exact}(\alpha) := f\Big(\underbrace{x_{\text{old}} - \alpha\, f_x(x_{\text{old}}, y_{\text{old}})}_{x(\alpha)}, \underbrace{y_{\text{old}} - \alpha\, f_y(x_{\text{old}}, y_{\text{old}})}_{y(\alpha)}\Big).$$

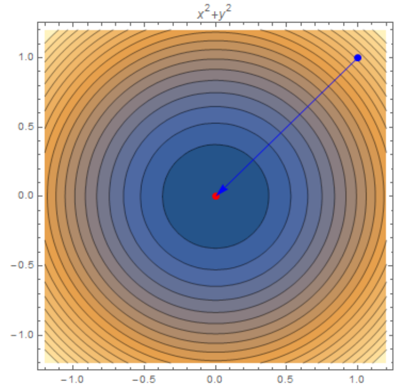
Group Problem 2



Track two steps of the steepest descent method applied to each function starting from the current guess $(x_{\text{old}}, y_{\text{old}}) = (1, 1)$, using (i) backtracking line-search, (ii) exact line-search.

Group Problem 2 (solution)

For the function $f(x, y) = x^2 + y^2$ (pictured above on the left), both line-searches locate the minimizer after one step, and, since the gradient at the minimizer is the zero-vector, don't move at subsequent steps.



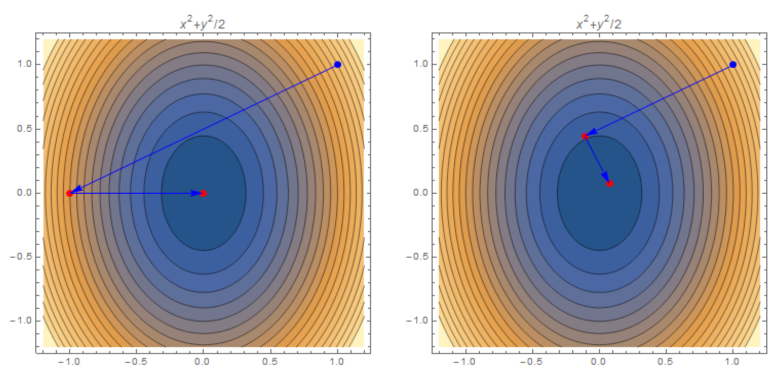
For the function $f(x, y) = x^2 + y^2/2$, the backtracking line-search makes it to the minimizer after two steps; moving from $(1, 1)$ to $(-1, 0)$ (using step-factor $\alpha = 1$) and from there to $(0, 0)$ (using step-factor $\alpha = \frac{1}{2}$). The exact line-search ends at $(2/27, 2/27)$ after two steps via the following calculations: From $(1, 1)$ we choose step-factor $\alpha = \frac{5}{9}$, which minimizes

$$\begin{aligned} \text{exact}(\alpha) &= f\Big(x_{\text{old}} - \alpha\, f_x(x_{\text{old}}, y_{\text{old}}),\ y_{\text{old}} - \alpha\, f_y(x_{\text{old}}, y_{\text{old}})\Big) \\ &= f(1 - 2\,\alpha,\ 1 - \alpha) = (1 - 2\,\alpha)^2 + (1 - \alpha)^2/2. \end{aligned}$$

The new guess after one step is thus $\left(1 - 2\,\frac{5}{9},\ 1 - \frac{5}{9}\right) = (-1/9, 4/9)$. From here, we choose step-factor $\alpha = \frac{5}{6}$ to minimize

$$\begin{aligned} \text{exact}(\alpha) &= f\Big(x_{\text{old}} - \alpha\, f_x(x_{\text{old}}, y_{\text{old}}),\ y_{\text{old}} - \alpha\, f_y(x_{\text{old}}, y_{\text{old}})\Big) \\ &= f(-1/9 + 2/9\,\alpha,\ 4/9 - 4/9\,\alpha) = (-1/9 + 2/9\,\alpha)^2 + (4/9 - 4/9\,\alpha)^2/2. \end{aligned}$$

The new guess after two steps is thus $(-1/9 + 10/54,\ 4/9 - 20/54) = (2/27, 2/27)$.



The good performance here of backtracking is a coincidence for this objective function and starting guess, and the exact line-search typically approaches the minimizer in fewer steps than backtracking.

See With Backtracking in Class8.nb.

2.2.1 Step-directions always perpendicular

You may have noticed in this example that the step-directions resulting from steepest-descent with an exact line-search are perpendicular to their predecessors. This turns out to happen in general.

To understand why, notice that to minimize the function

$$\operatorname{exact}(\alpha) := f\Big(\underbrace{x_{\text{old}} - \alpha\, f_x(x_{\text{old}}, y_{\text{old}})}_{x(\alpha)},\ \underbrace{y_{\text{old}} - \alpha\, f_y(x_{\text{old}}, y_{\text{old}})}_{y(\alpha)}\Big)$$

with respect to α, we can set its derivative with respect to α equal to zero:

$$\begin{aligned} 0 &= \operatorname{exact}'(\alpha) \\ &= f_x\big(x(\alpha),\, y(\alpha)\big) \cdot x'(\alpha) + f_y\big(x(\alpha),\, y(\alpha)\big) \cdot y'(\alpha) \\ &= f_x\big(x(\alpha),\, y(\alpha)\big)\,\big(-f_x\,(x_{\text{old}}, y_{\text{old}})\,\big) + f_y\big(x(\alpha),\, y(\alpha)\big)\,\big(-f_y\,(x_{\text{old}}, y_{\text{old}})\,\big) \end{aligned}$$

where we have applied the two-variable chain rule since α appears in both the $x(\alpha)$ and $y(\alpha)$ arguments of the two-variable function f. Since we know that the minimizing α solves this equation, we can substitute

$$(x_{\text{new}}, y_{\text{new}}) = (x(\alpha),\, y(\alpha))$$

to get

$$0 = f_x\,(x_{\text{new}}, y_{\text{new}})\ (-f_x\,(x_{\text{old}}, y_{\text{old}})) + f_y\,(x_{\text{new}}, y_{\text{new}})\ (-f_y\,(x_{\text{old}}, y_{\text{old}}))$$

which implies that the dot product $\nabla f\,(x_{\text{new}}, y_{\text{new}}) \cdot -\nabla f\,(x_{\text{old}}, y_{\text{old}})$ is zero. This ensures that the consecutive step-directions are perpendicular.

Group Problem 3

Explain the perpendicularity of consecutive step-directions in steepest descent with an exact line-search conceptually by describing why the incoming line of search must be parallel to the contour of f at the new guess.

Group Problem 3 (solution)

We know that a minimizer of f along any fixed direction will be on the line through that direction at a point where that line is parallel to a contour of f. We can see this by considering the alternative: If the line instead crossed the contour, then moving along the line either forward or backward from that point would definitely decrease the value of f. This contradicts the fact that an exact line-search stops at a minimizer along that direction. Therefore, we know that the old step-direction is parallel to the contour of f through the new guess.

It is a fact that the negative gradient $-\nabla f(x, y)$ at any point (x, y) is perpendicular to the contour of f through (x, y). Thus, we know that the new step-direction from the new guess is perpendicular to the contour of f through the new guess. Since the old step-direction is parallel to this same contour, the new step-direction must be perpendicular to the old step-direction.