- Unconstrained minimization

$$\min. \ f(x)$$

f convex, twice continuously differentiable (hence dom f open)
assume $p^* = \inf_x f(x)$ is attained and finite

Solve optimality condition : $\nabla f(x^*) = 0$

1. initial point and sub-level set

$x^{(0)} \in \text{dom} f$

$S = \{x \mid f(x) \le f(x^{(0)})\}$ is closed $\Rightarrow$ { hard to verify, except when all sublevel sets are closed:
equivalent to the condition that : epi(f) is closed
true if dom $f = \mathbb{R}^n$
true if $f(x) \to \infty$ as $x \to$ boundary (dom f)

eg. $f(x) = \log \sum \exp(a_i^T x + b_i)$
dom $f = \mathbb{R}^n$

eg. $f(x) = -\sum \log(b_i - a_i^T x)$
domain is an open polyhedron
$b_i - a_i^T x \to 0$ as $x \to$ boundary

- Strong convexity and implications

f is strongly convex on S if there exists an $m > 0$ such that
$\nabla^2 f(x) \succeq mI$ for all $x \in S$

$$f(y) = f(x) + \nabla f(x)^T (y-x) + \tfrac{1}{2}(y-x)^T \nabla^2 f(z)(y-x) \quad \text{for some } z \in [x, y]$$
$$f(y) \ge f(x) + \nabla f(x)^T (y-x) + m/2 \, \|y-x\|_2^2$$

$$\begin{aligned} f(y) &\ge f(x) + \nabla f(x)^T (y-x) + m/2 \, \|y-x\|_2^2 \\ &\ge f(x) + \nabla f(x)^T (\tilde{y}-x) + m/2 \, \|\tilde{y}-x\|_2^2 \\ &= f(x) - 1/2m \, \|\nabla f(x)\|_2^2 \end{aligned}$$

$\nabla_y = \nabla f(x) + m\|y-x\| := 0$

$\tilde{y} = \dfrac{-\nabla f(x)}{m} + x$

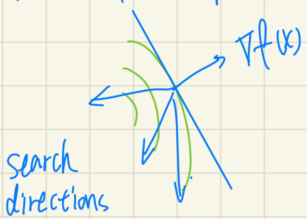
$f(x) - p^* \le 1/2m \, \|\nabla f(x)\|_2^2$ (Stopping criterion)

- Descent methods

$$x^{(k+1)} = x^{(k)} + t^{(k)} \cdot \Delta x^{(k)} \quad \text{with } f(x^{(k+1)}) < f(x^{(k)})$$

Δx is the search direction ; t is the step size

from convexity . $f(x^{(k+1)}) < f(x^{(k)})$ implies that $\nabla f(x)^T \Delta x < 0$



- Line search

1. exact line search

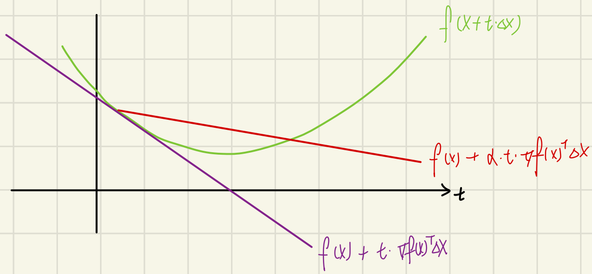
$$t = \arg\min_{t>0} f(x + t \cdot \Delta x)$$

2. backtracking line search (with parameter $\alpha \in (0, \frac{1}{2})$ $\beta \in (0,1)$)

Starting at $t=1$, repeat $t := \beta t$ until

$$f(x + t\Delta x) < f(x) + \alpha \cdot t \cdot \nabla f(x)^T \Delta x$$

$f(x + t\Delta x) < f(x) + t \nabla f(x)^T \Delta x$ will never by convexity



- Gradient descent method

$\Delta x = -\nabla f(x)$

giving $x^{(0)} \in \operatorname{dom} f$

repeat {

$\Delta x := -\nabla f(x)$

line search to chose t

$x := x + t\Delta x$

}

Stopping criterion : $\|\nabla f(x)\|_2 \le \epsilon$

Convergence result : for strongly convex f

$$f(x^{(k)}) - p^* \le c^k \left(f(x^{(0)}) - p^* \right)$$

$c \in [0,1)$ depends on m, $x^{(0)}$ and line search type

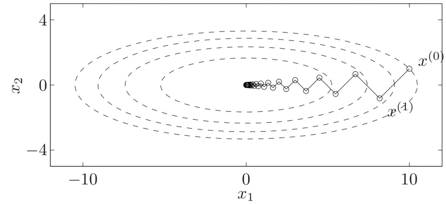


Figure 9.2 Some contour lines of the function $f(x) = (1/2)(x_1^2 + 10x_2^2)$. The condition number of the sublevel sets, which are ellipsoids, is exactly 10. The figure shows the iterates of the gradient method with exact line search, started at $x^{(0)} = (10, 1)$.

- Steepest descent method

1. normalized steepest descent direction

$$\Delta x_{nsd} = \underset{v}{\mathrm{argmin}} \{ \nabla f(x)^T v \mid \|v\| = 1 \}$$ directional derivative along $x + tv$

$f(x+v) \approx f(x) + \nabla f(x)^T v$ for small v

Δx_{nsd} is unit-norm step with most negative directional derivative

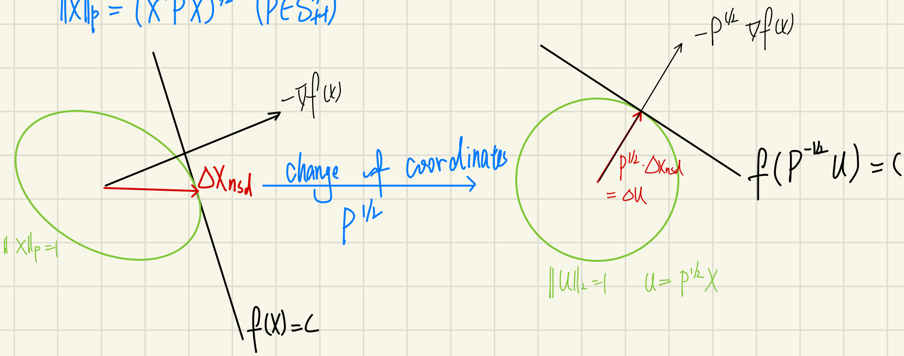
2. unnormalized steepest descent direction

$$\Delta x_{sd} = \|\nabla f(x)\|_* \cdot \Delta x_{nsd}$$ $\|\nabla f(x)\|_* = \sup_z \{ z^T \nabla f(x) \mid \|z\| = 1 \}$

$\nabla f(x)^T \Delta x_{sd} = -\|\nabla f(x)\|_2$

Gradient descent is steepest descent in euclidean norm eg. Quadratic norm

$$\|x\|_P = (x^T P x)^{1/2} \quad (P \in S^n_{++})$$



$$\Delta u = -P^{1/2} \cdot \nabla f(P^{-1/2} u) = -P^{1/2} \cdot \nabla f(x)$$

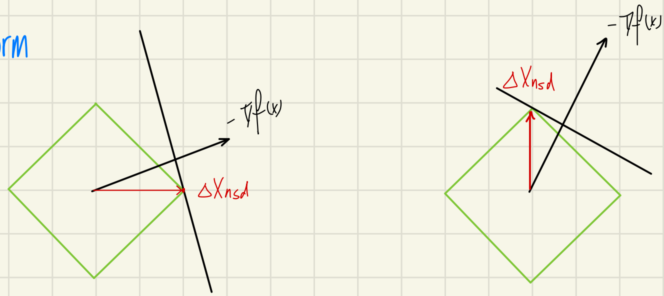
$$\Delta u_{nsd} = -P^{1/2} \cdot \nabla f(x) \,/\, \|-P^{1/2} \cdot \nabla f(x)\|_2$$
$$= -P^{1/2} \cdot \nabla f(x) \,/\, (\nabla f(x)^T P \nabla f(x))^{1/2}$$

$$\Delta u_{sd} = -P^{1/2} \cdot \nabla f(x)$$

$$\Delta x_{nsd} = P^{1/2} \cdot \Delta u_{nsd}$$
$$= -P^{-1} \cdot \nabla f(x) \,/\, (\nabla f(x)^T P \nabla f(x))^{1/2}$$

$$\Delta x_{sd} = -P \cdot \nabla f(x)$$

eg. L_1 norm



$$\Delta x_{nsd} = -\text{sign}\left(\frac{\partial f}{\partial x_i}\right) \cdot e_i \qquad i = \arg\max_i \left|\frac{\partial f}{\partial x_i}\right|$$

Steepest descent in L^1: update 1 component each step

3. Choice of norm for steepest descent

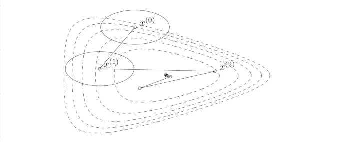


Figure 9.11 Steepest descent method with a quadratic norm $\|\cdot\|_{P_1}$. The ellipses are the boundaries of the norm balls $\{x \mid \|x - x^{(k)}\|_{P_1} \le 1\}$ at $x^{(0)}$ and $x^{(1)}$.

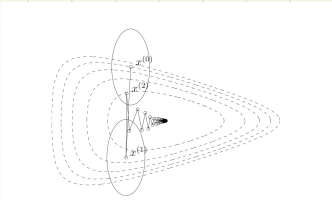


Figure 9.12 Steepest descent method, with quadratic norm $\|\cdot\|_{P_2}$.

want the norm to be consistent with the geometry of the sublevel sets

$$f(x) \approx f(x^*) + \underbrace{\nabla f(x^*)^T (x-x^*)}_{0} + (x-x^*)^T \nabla^2 f(x^*) (x-x^*)$$

sublevel sets are ellipsoids (f is nearly quadratic) near the optimal point

{Steepest descent in the norm induced by the hessian $\Rightarrow$ Newton's method}

- Newton step

$$\Delta x_{nt} = -\nabla^2 f(x)^{-1} \cdot \nabla f(x)$$

1. interpretation 1.

   $x + \Delta x_{nt}$ minimizes the second-order approximation

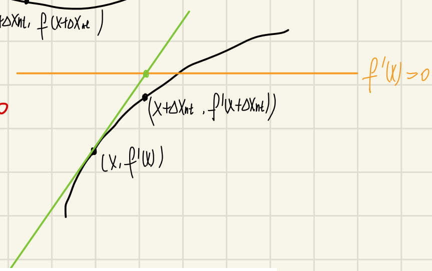
$$\hat{f}(x+v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v$$

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

2. interpretation 2

   $x + \Delta x_{nt}$ solves the optimality condition

$$\nabla f(x+v) \approx \nabla f(x) + \nabla^2 f(x) \cdot v := 0$$



3. interpretation 3.

   Steepest descent in local hessian norm

$$\|u\|_{\nabla^2 f(x)} = (u^T \nabla^2 f(x) u)^{1/2}$$

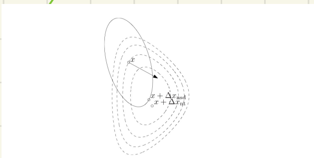


Figure 9.17 The dashed lines are level curves of a convex function. The ellipsoid shown (with solid line) is $\{x+v \mid v^T \nabla^2 f(x) v \le 1\}$. The arrow shows $-\nabla f(x)$, the gradient descent direction. The Newton step Δx_{nt} is the steepest descent direction in the norm $\|\cdot\|_{\nabla^2 f(x)}$. The figure also shows Δx_{nsd}, the normalized steepest descent direction for the same norm.

4. Affine invariance of newton step.

   suppose $\bar{f}(y) = f(Ty)$ $\quad T \in S^n_{++}$

$$\nabla \bar{f}(y) = T^T \nabla f(y) \qquad \nabla^2 \bar{f}(y) = T^T \nabla^2 f(x) T$$

$\nabla \tilde{f}(y) = T \cdot f(x)$ gradient method needs proper coordinate

$$\Delta y_{nt} = -(T^T \nabla^2 f(x) T)^{-1} \cdot T^T \nabla f(x)$$

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

$$= T^{-1} \Delta x_{nt}$$

$$x + \Delta x_{nt} = T(y + \Delta y_{nt})$$

newton step is independent of choice of coordinates

5. Newton decrement

$$\lambda(x) = \left(\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)\right)^{1/2}$$ the hessian-norm of gradient

1° gives a estimate of decrease

$$\Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$$

$$\begin{aligned} f(x) - \inf_y \hat{f}(y) &= f(x) - \hat{f}(x + \Delta x_{nt}) \\ &= -\nabla f(x)^T \Delta x_{nt} - \tfrac{1}{2} \Delta x_{nt}^T \nabla^2 f(x) \Delta x_{nt} \qquad = \tfrac{1}{2} \Delta x_{nt}^T \nabla^2 f(x) \Delta x_{nt} \\ &= \tfrac{1}{2} \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) \quad = \tfrac{1}{2} \lambda(x)^2 \end{aligned}$$

· Newton's method

given a starting point $x \in \mathrm{dom} f$, tolerance $\epsilon > 0$

repeat {

compute the newton step and decrement

$$\Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$$
$$\lambda^2 = \nabla f(x) \nabla^2 f(x)^{-1} \nabla f(x)$$

stop if $\lambda^2/2 \le \epsilon$

line search to choose step size t

$x := x + t \cdot \Delta x_{nt}$

}

$$\left.\begin{tikzcd} \min.\ f(x) \arrow[rr, leftrightarrow, "\text{change coordinates}"] \arrow[d, leftrightarrow, "\text{newton}"] & & \min.\ f(Ty) \arrow[d, leftrightarrow, "\text{newton}"] \\ x^{(1)} \cdots x^{(k)} \arrow[rr, leftrightarrow, "\text{change coordinates}"'] & & y^{(1)} \cdots y^{(k)} \end{tikzcd}\right\}$$

affine - invariant

commute diagram

· Newton method convergence analysis

Assume f is strongly convex on S with constant $m > 0$ s.t. $\nabla^2 f(x) \succeq mI$.

$\nabla^2 f$ is Lipschitz continuous on S, with constant $L > 0$

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \le L \|x - y\|_2$$ (affine transformation changes norm !!!)

L is a constraint on third derivative ; L measures how well f is approximated by a quadratic function

(L can be taken 0 for a quadratic function ; newton method works perfectly (1 step) for quadratic function)

(when the 3rd derivative is small, i.e. the hessian is slowly changing, newton method works well) there exist constant $\eta \in (0, m^2/L)$, $\gamma > 0$ such that

if $\|\nabla f(x)\|_2 \ge \eta$, then $f(x^{(k)}) - f(x^{(k+1)}) \ge \gamma$ when gradient large, guaranteed to decrease by a certain constant

if $\|\nabla f(x)\|_2 < \eta$ then

$$\frac{L}{2m^2}\|\nabla f(x^{(k+1)})\|_2 \le \left(\frac{L}{2m^2}\|\nabla f(x^{(k)})\|_2\right)^2$$

1. damped Newton phase ($\|\nabla f(x)\|_2 \ge \eta$)
   - most iterations require backtracking steps
   - function value decreases by at least γ each step
   - if $p^* > -\infty$, the phase takes at most $(f(x^{(0)}) - p^*)/\gamma$ iterations

2. quadratically convergent phase ($\|\nabla f(x)\|_2 < \eta$)
   - all iterations use step size $t=1$
   - $\|\nabla f(x)\|_2$ converges to zero quadratically: if $\|\nabla f(x^{(k)})\|_2 < \eta$, then

$$\frac{L}{2m^2}\|\nabla f(x^{l})\|_2 \le \left(\frac{L}{2m^2}\|\nabla f(x^{(k)})\|_2\right)^{2^{l-k}} \le \underline{(1/2)^{2^{l-k}}} \qquad (l \ge k)$$

number of accurate digits doubles every step.

number of iterations until $f(x) - p^* \le \epsilon$ is bounded above by:

$$\underbrace{\frac{f(x^{(0)}) - p^*}{\gamma}}_{\substack{\text{first stage;} \\ \text{depends on how suboptimal } x^{(0)} \text{ is}}} + \log_2 \log_2 (\epsilon_0/\epsilon)$$

eg. $\mathbb{R}^2$

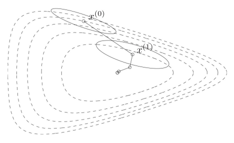


Figure 9.19 Newton's method for the problem in $\mathbf{R}^2$, with objective f given in (9.20), and backtracking line search parameters $\alpha = 0.1$, $\beta = 0.7$. Also shown are the ellipsoids $\{x \mid \|x - x^{(k)}\|_{\nabla^2 f(x^{(k)})} \le 1\}$ at the first two iterates.

damped newton phase | quadratically convergent

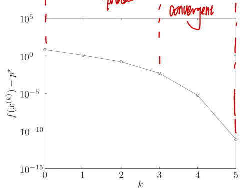


Figure 9.20 Error versus iteration k of Newton's method for the problem in $\mathbf{R}^2$. Convergence to a very high accuracy is achieved in five iterations.

each newton step computes $Hv = -g$