# Unconstrained minimization problems

- Strong convexity and implications

**Condition number of sublevel sets**

We define the *width* of a convex set $C \subseteq \mathbb{R}^n$, in the direction $\mathbf{q}$, where $\|\mathbf{q}\|_2 = 1$, as

$$W(C, \mathbf{q}) = \sup_{\mathbf{z} \in C} \mathbf{q}^T \mathbf{z} - \inf_{\mathbf{z} \in C} \mathbf{q}^T \mathbf{z}.$$

The *minimum width* and maximum width of $C$ are given by

$$W_{\min} = \inf_{\|\mathbf{q}\|_2 = 1} W(C, \mathbf{q}), \quad W_{\max} = \sup_{\|\mathbf{q}\|_2 = 1} W(C, \mathbf{q}).$$

The *condition number* of the convex set $C$ is defined as

$$\mathbf{cond}(C) = \frac{W_{\max}^2}{W_{\min}^2},$$

*i.e.*, the square of the ratio of its maximum width to its minimum width. The condition number of $C$ gives a measure of its *anisotropy* or *eccentricity*.

# Unconstrained minimization problems

- Strong convexity and implications

Example. *Condition number of an ellipsoid.* Let $\mathcal{E}$ be the ellipsoid

$$\mathcal{E} = \{\mathbf{x} | (\mathbf{x} - \mathbf{x}_0)^T \mathbf{A}^{-1} (\mathbf{x} - \mathbf{x}_0) \leq 1\},$$

where $\mathbf{A} \in \mathbb{S}^n_{++}$. The width of $\mathcal{E}$ in the direction $\mathbf{q}$ is

$$\sup_{\mathbf{z} \in \mathcal{E}} \mathbf{q}^T \mathbf{z} - \inf_{\mathbf{z} \in \mathcal{E}} \mathbf{q}^T \mathbf{z} = (\|\mathbf{A}^{1/2}\mathbf{q}\|_2 + \mathbf{q}^T \mathbf{x}_0) - (-\|\mathbf{A}^{1/2}\mathbf{q}\|_2 + \mathbf{q}^T \mathbf{x}_0) = 2\|\mathbf{A}^{1/2}\mathbf{q}\|_2.$$

It follows that its minimum and maximum width are

$$W_{\min} = 2\lambda_{\min}(\mathbf{A})^{1/2}, \quad W_{\max} = 2\lambda_{\max}(\mathbf{A})^{1/2}.$$

and its condition number is

$$\mathbf{cond}(\mathcal{E}) = \frac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{A})} = \kappa(\mathbf{A}).$$

# Unconstrained minimization problems

- Strong convexity and implications

Suppose $f$ satisfies $m\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq M\mathbf{I}$ for all $\mathbf{x} \in S$. We will derive a bound on the condition number of the $\alpha$-sublevel $C_\alpha = \{\mathbf{x} | f(\mathbf{x}) \leq \alpha\}$, where $p^* < \alpha \leq f(\mathbf{x}^{(0)})$. We have

$$p* + (M/2)\|\mathbf{y} - \mathbf{x}^*\|_2^2 \geq f(\mathbf{y}) \geq p^* + (m/2)\|\mathbf{y} - \mathbf{x}^*\|_2^2.$$

This implies that $B_{inner} \subseteq C_\alpha \subseteq B_{outer}$ where

$$\begin{aligned} B_{inner} &= \{\mathbf{y} | \|\mathbf{y} - \mathbf{x}^*\|_2 \leq (2(\alpha - p^*)/M)^{1/2}\}, \\ B_{outer} &= \{\mathbf{y} | \|\mathbf{y} - \mathbf{x}^*\|_2 \leq (2(\alpha - p^*)/m)^{1/2}\}. \end{aligned}$$

The ratio of the radii squared gives an upper bound on the condition number of $C_\alpha$:

$$\mathbf{cond}(C_\alpha) \leq \frac{M}{m}.$$

# Descent methods

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t^{(k)}\Delta\mathbf{x}^{(k)},$$

search direction

step size

where $t^{(k)} > 0$ (except when $\mathbf{x}^{(k)}$ is optimal).

*Descent methods* mean that

$$f(\mathbf{x}^{(k+1)}) < f(\mathbf{x}^{(k)}),$$

except when $\mathbf{x}^{(k)}$ is optimal.

From convexity we know that $\nabla f(\mathbf{x}^{(k)})^T(\mathbf{y} - \mathbf{x}^{(k)}) \geq 0$ implies $f(\mathbf{y}) \geq f(\mathbf{x}^{(k)})$, so the search direction in a descent method must satisfy

$$\nabla f(\mathbf{x}^{(k)})^T \Delta\mathbf{x}^{(k)} < 0,$$

*i.e.*, it must make an acute angle with the negative gradient. We call such a direction a *descent direction* (for $f$, at $\mathbf{x}^{(k)}$).

# Descent methods

**Algorithm 5.1** *General descent method.*

**given** a starting point $\mathbf{x} \in \mathbf{dom} f$.

**repeat**

1. Determine a descent direction $\Delta\mathbf{x}$.
2. Line search. Choose a step size $t > 0$.
3. Update. $\mathbf{x} := \mathbf{x} + t\Delta\mathbf{x}$.

**until** stopping criterion is satisfied.

The stopping criterion is often of the form $\|\nabla f(\mathbf{x})\|_2 \leq \eta$, where $\eta$ is small and positive.

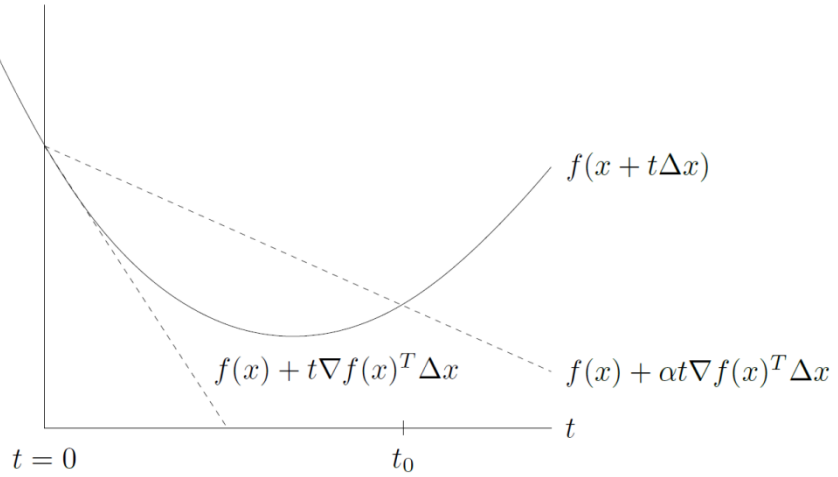
# Descent methods

- Exact line search

$$t = \operatorname*{argmin}_{s \geq 0} f(\mathbf{x} + s\Delta\mathbf{x}).$$

- Backtracking line search

**Algorithm 9.2** *Backtracking line search.*
**given** a descent direction $\Delta\mathbf{x}$ for $f$ at $\mathbf{x} \in \mathbf{dom} f, \alpha \in (0, 0.5), \beta \in (0, 1)$.
$t := 1$.
**while** $f(\mathbf{x} + t\Delta\mathbf{x}) > f(\mathbf{x}) + \alpha t \nabla f(\mathbf{x})^T \Delta\mathbf{x}, t := \beta t$.



# Descent methods

- Backtracking line search

The line search is called *backtracking* because it starts with unit step size and then reduces it by the factor $\beta$ until the stopping condition

$$f(\mathbf{x} + t\Delta\mathbf{x}) \leq f(\mathbf{x}) + \alpha t \nabla f(\mathbf{x})^T \Delta\mathbf{x}$$

holds. Since $\Delta\mathbf{x}$ is a descent direction, we have $\nabla f(\mathbf{x})^T \Delta\mathbf{x} < 0$, so for small enough $t$ we have

$$f(\mathbf{x} + t\Delta\mathbf{x}) \approx f(\mathbf{x}) + t\nabla f(\mathbf{x})^T \Delta\mathbf{x} < f(\mathbf{x}) + \alpha t \nabla f(\mathbf{x})^T \Delta\mathbf{x},$$

which shows that the backtracking line search eventually terminates.

# Descent methods

- Gradient descent method

A natural choice for the search direction is the negative gradient

$$\Delta \mathbf{x} = -\nabla f(\mathbf{x}).$$

The resulting algorithm is called the *gradient algorithm* or *gradient descent method.*

**Algorithm 5.3** *Gradient descent method.*

**given** a starting point $\mathbf{x} \in \mathbf{dom} f$.

**repeat**

1. $\Delta \mathbf{x} := -\nabla f(\mathbf{x})$.
2. Line search. Choose step size $t$ via exact or backtracking line search.
3. Update. $\mathbf{x} := \mathbf{x} + t\Delta \mathbf{x}$.

**until** stopping criterion is satisfied.

The stopping criterion is usually of the form $\|\nabla f(\mathbf{x})\|_2 \leq \eta$, where $\eta$ is small and positive.

# Descent methods

- Convergence analysis

Lighter notations: $\mathbf{x}^+ = \mathbf{x} + t\Delta\mathbf{x}$ for $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t^{(k)}\Delta\mathbf{x}^{(k)}$, where $\Delta\mathbf{x} = -\nabla f(\mathbf{x})$. We assume $f$ is strongly convex on $S$, so there are positive constants $m$ and $M$ such that $m\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq M\mathbf{I}$ for all $\mathbf{x} \in S$. Define the function $\tilde{f} : \mathbb{R} \to \mathbb{R}$ by $\tilde{f}(t) = f(\mathbf{x} - t\nabla f(\mathbf{x}))$. From the inequality (6), with $\mathbf{y} = \mathbf{x} - t\nabla f(\mathbf{x})$, we obtain a quadratic upper bound on $\tilde{f}$:

$$\tilde{f}(t) \leq f(\mathbf{x}) - t\|\nabla f(\mathbf{x})\|_2^2 + \frac{Mt^2}{2}\|\nabla f(\mathbf{x})\|_2^2. \tag{8}$$

# Descent methods

- Convergence analysis

**Analysis for exact line search**

Minimize over $t$ both sides of the inequality (8). We have

$$f(\mathbf{x}^+) = \tilde{f}(t_{exact}) \leq f(\mathbf{x}) - \frac{1}{2M}\|\nabla(f(\mathbf{x}))\|_2^2.$$

Subtracting $p^*$ from both sides, we get

$$f(\mathbf{x}^+) - p^* \leq f(\mathbf{x}) - p^* - \frac{1}{2M}\|\nabla f(\mathbf{x})\|_2^2.$$

We combine this with $\|\nabla f(\mathbf{x})\|_2^2 \leq 2m(f(\mathbf{x}) - p^*)$ (from (2)) to conclude

$$f(\mathbf{x}^+) - p^* \leq (1 - m/M)(f(\mathbf{x}) - p^*).$$

Applying this inequality recursively, we find that

linear convergence!
$M/m$: condition number

$$f(\mathbf{x}^{(k)}) - p^* \leq c^k(f(\mathbf{x}^{(0)}) - p^*), \text{ where } c = 1 - m/M < 1. \quad (9)$$

# Descent methods

- Convergence analysis

**Analysis for backtracking line search** The backtracking exit condition,

$$\tilde{f}(t) \leq f(\mathbf{x}) - \alpha t \|\nabla f(\mathbf{x})\|_2^2,$$

is satisfied whenever $0 \leq t \leq 1/M$. First note that $0 \leq t \leq 1/M \implies -t + \frac{Mt^2}{2} \leq -t/2$. Using this result and the bound (8), we have, for $0 \leq t \leq 1/M$,

$$\begin{aligned}\tilde{f}(t) \leq & f(\mathbf{x}) - t\|\nabla f(\mathbf{x})\|_2^2 + \frac{Mt^2}{2}\|\nabla f(\mathbf{x})\|_2^2 \\ \leq & f(\mathbf{x}) - (t/2)\|\nabla f(\mathbf{x})\|_2^2 \leq f(\mathbf{x}) - \alpha t\|\nabla f(\mathbf{x})\|_2^2\end{aligned}$$

since $\alpha < 1/2$. Therefore the backtracking line search terminates either with $t = 1$ or with a value $t \geq \beta/M$. In the first case we have

$$f(\mathbf{x}^+) \leq f(\mathbf{x}) - \alpha\|\nabla f(\mathbf{x})\|_2^2,$$

and in the second case we have

$$f(\mathbf{x}^+) \leq f(\mathbf{x}) - (\beta\alpha/M)\|\nabla f(\mathbf{x})\|_2^2.$$

# Descent methods

- Convergence analysis

  Putting these together, we always have

$$f(\mathbf{x}^+) \leq f(\mathbf{x}) - \min\{\alpha, (\beta\alpha/M)\}\|\nabla f(\mathbf{x})\|_2^2.$$

We subtract $p^*$ from both sides to get

$$f(\mathbf{x}^+) - p^* \leq f(\mathbf{x}) - p^* - \min\{\alpha, (\beta\alpha/M)\}\|\nabla f(\mathbf{x})\|_2^2,$$

and combine this with $\|\nabla f(\mathbf{x})\|_2^2 \geq 2m(f(\mathbf{x}) - p^*)$ to obtain

$$f(\mathbf{x}^+) - p^* \leq (1 - \min\{2m\alpha, 2\beta\alpha m/M\})(f(\mathbf{x}) - p^*).$$

From this we conclude

$$f(\mathbf{x}^{(k)}) - p^* \leq c^k(f(\mathbf{x}^{(0)}) - p^*),$$

where $c = 1 - \min\{2m\alpha, 2\beta\alpha m/M\} < 1$.

# Descent methods

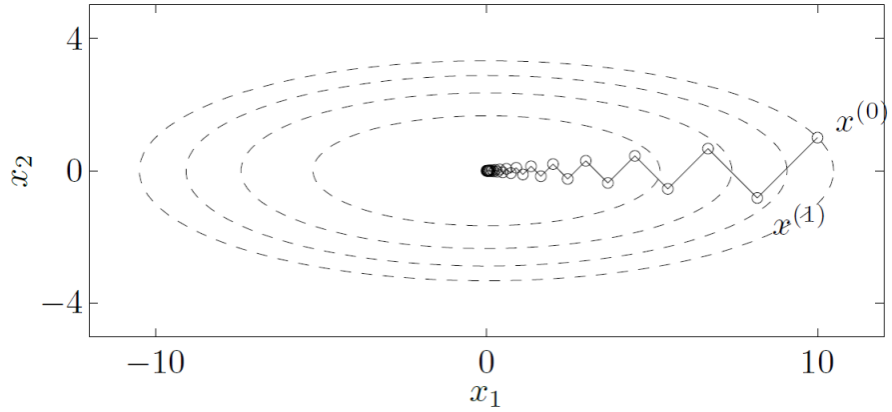
- Examples

**A quadratic problem in $\mathbb{R}^2$.**

$$f(\mathbf{x}) = \frac{1}{2}(x_1^2 + \gamma x_2^2),$$

where $\gamma > 0$. Clearly, the optimal point is $\mathbf{x}^* = \mathbf{0}$, and the optimal value is 0. The Hessian of $f$ is constant, and has eigenvalues 1 and $\gamma$, so the condition numbers of the sublevel sets of $f$ are all exactly

$$\frac{\max\{1,\gamma\}}{\min\{1,\gamma\}} = \max\{\gamma, 1/\gamma\}.$$



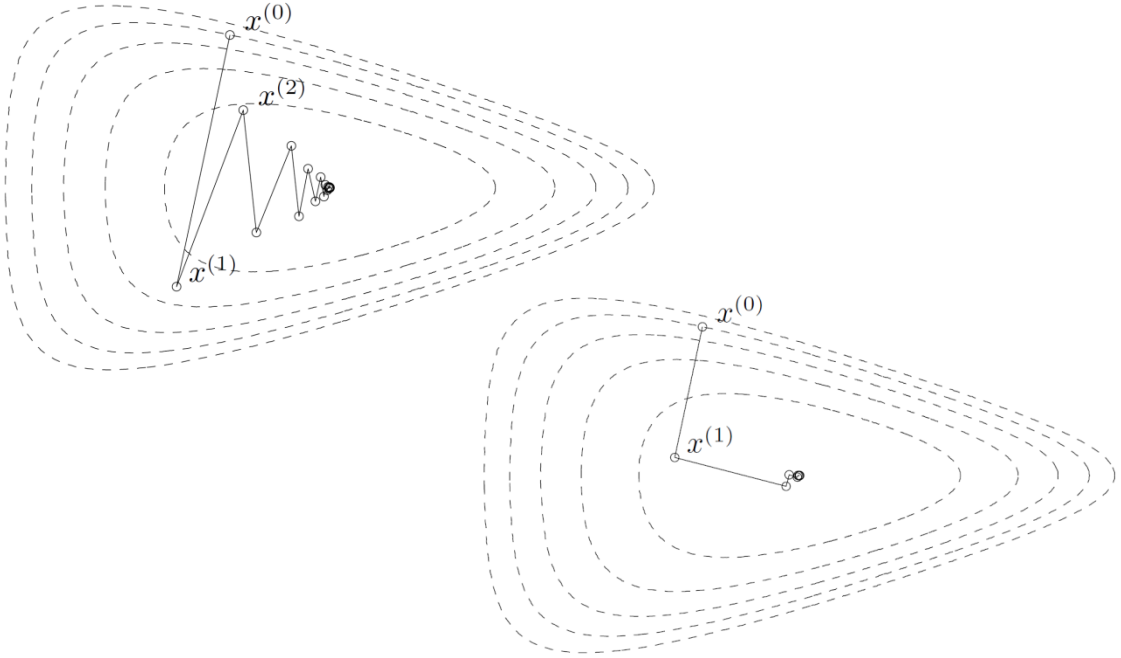
# Descent methods

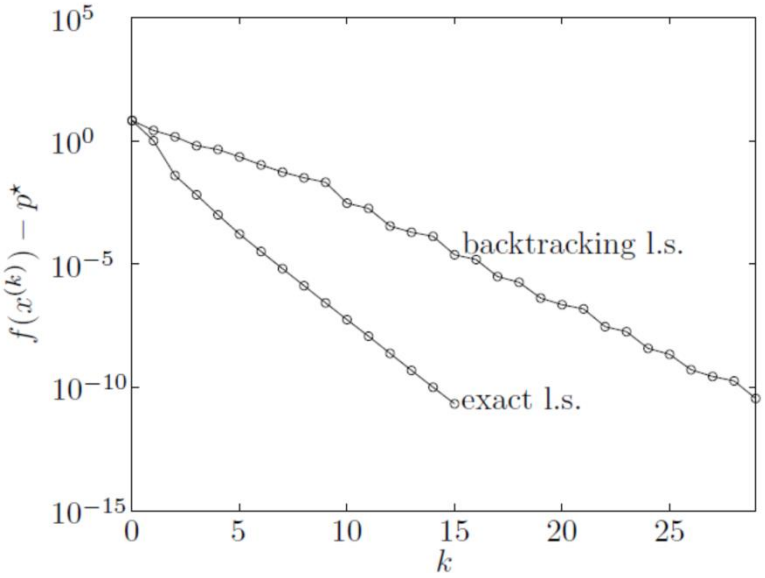
- Examples

**A nonquadratic problem in $\mathbb{R}^2$.**

$$f(x_1, x_2) = e^{x_1+3x_2-0.1} + e^{x_1-3x_2-0.1} + e^{-x_1-0.1}.$$

We apply the gradient method with a backtracking line search, with $\alpha = 0.1$, $\beta = 0.7$.



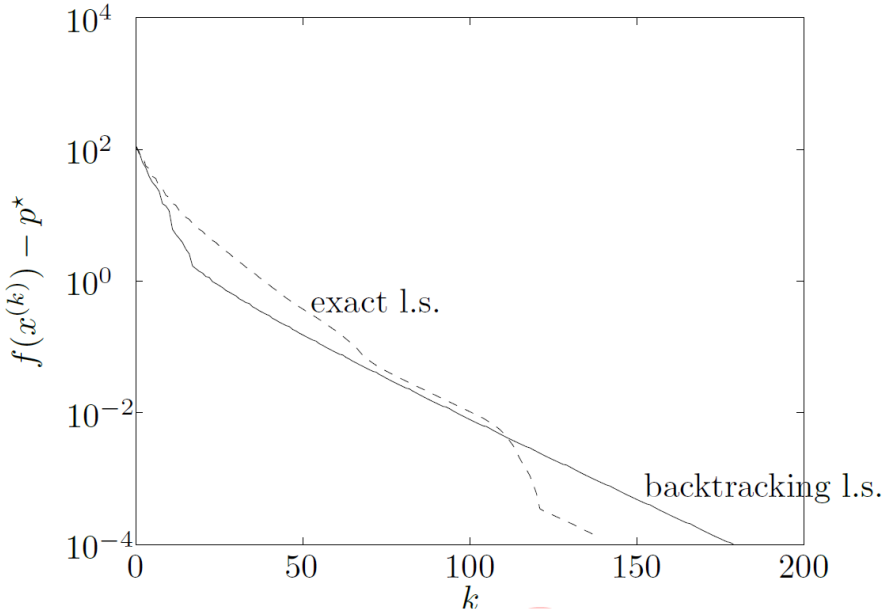
# Descent methods

- Examples

**A problem in $\mathbb{R}^{100}$.**

$$f(\mathbf{x}) = \mathbf{c}^T\mathbf{x} - \sum_{i=1}^{m} \log(b_i - \mathbf{a}_i^T\mathbf{x}),$$

with $m = 500$ terms and $n = 100$ variables. $\alpha = 0.1$ and $\beta = 0.5$



# Descent methods

- Steepest descent method

The first-order Taylor approximation of $f(\mathbf{x}+\mathbf{v})$ around $\mathbf{x}$ is

$$f(\mathbf{x}+\mathbf{v}) \approx \hat{f}(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T \mathbf{v}.$$

directional derivative

Normalized steepest descent direction (with respect to the norm $\|\cdot\|$):

$$\Delta \mathbf{x}_{nsd} = \underset{\mathbf{v}}{\operatorname{argmin}}\{\nabla f(\mathbf{x})^T \mathbf{v} | \|\mathbf{v}\| = 1\}.$$

It is also convenient to consider a steepest descent step $\Delta \mathbf{x}_{sd}$ that is unnormalized:

$$\Delta \mathbf{x}_{sd} = \|\nabla f(\mathbf{x})\|_* \Delta \mathbf{x}_{nsd},$$

$\Delta \mathbf{x}_{sd} = -\nabla f(\mathbf{x})$ if $\|\cdot\|$ is the Euclidean norm.

where $\|\cdot\|_*$ denotes the dual norm. Note that for the steepest descent step, we have

$$\nabla f(\mathbf{x})^T \Delta \mathbf{x}_{sd} = \|\nabla f(\mathbf{x})\|_* \nabla f(\mathbf{x})^T \Delta \mathbf{x}_{nsd} = -\|\nabla f(\mathbf{x})\|_*^2.$$

# Descent methods

- Steepest descent method

**Algorithm 5.4** *Steepest descent method.*

**given** a starting point $\mathbf{x} \in \mathbf{dom} f$.

**repeat**

1. Compute steepest descent direction $\Delta\mathbf{x}_{sd}$.
2. Line search. Choose $t$ via backtracking or exact line search.
3. Update. $\mathbf{x} := \mathbf{x} + t\Delta\mathbf{x}_{sd}$.

**until** stopping criterion is satisfied.

Scale factors in the descent direction have no effect. So the unnormalized direction can be used.

# Descent methods

- Steepest descent for Euclidean and quadratic norms

**Steepest descent for Euclidean norm**

If we take the norm $\|\cdot\|$ to be the Euclidean norm we find that the steepest descent direction is simply the negative gradient, *i.e.*, $\Delta\mathbf{x}_{sd} = -\nabla f(\mathbf{x})$. The steepest descent method for the Euclidean norm coincides with the gradient descent method.

# Descent methods

- Steepest descent for Euclidean and quadratic norms

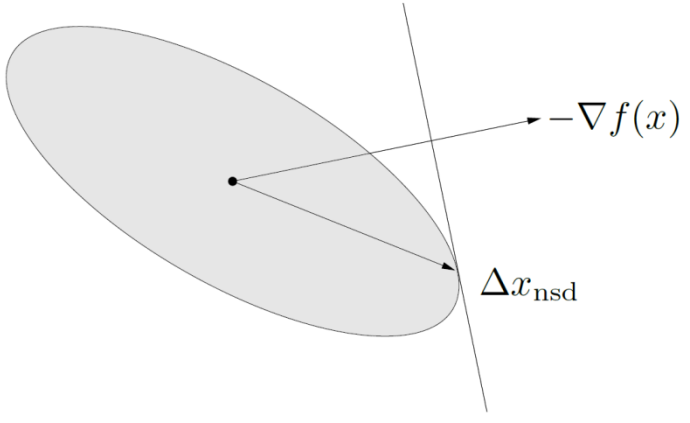
**Steepest descent for quadratic norm**

We consider the quadratic norm $\|\mathbf{z}\|_{\mathbf{P}} = (\mathbf{z}^T\mathbf{P}\mathbf{z})^{1/2} = \|\mathbf{P}^{1/2}\mathbf{z}\|_2$, where $\mathbf{P} \in \mathbb{S}^n_{++}$. The normalized steepest descent direction is given by

$$\Delta\mathbf{x}_{nsd} = -(\nabla f(\mathbf{x})^T\mathbf{P}^{-1}\nabla f(\mathbf{x}))^{-1/2}\mathbf{P}^{-1}\nabla f(\mathbf{x}).$$

The dual norm is given by $\|\mathbf{z}\|_* = \|\mathbf{P}^{-1/2}\mathbf{z}\|_2$, so the steepest descent step with respect to $\|\cdot\|_{\mathbf{P}}$ is given by

$$\Delta\mathbf{x}_{nsd} = -\mathbf{P}^{-1}\nabla f(\mathbf{x}).$$



# Descent methods

- Steepest descent for Euclidean and quadratic norms

**Interpretation via change of coordinates**

Define $\bar{\mathbf{u}} = \mathbf{P}^{1/2}\mathbf{u}$, so we have $\|\mathbf{u}\|_{\mathbf{P}} = \|\bar{\mathbf{u}}\|_2$. The equivalent problem of minimizing

$$\bar{f}(\bar{\mathbf{u}}) = f(\mathbf{P}^{-1/2}\bar{\mathbf{u}}) = f(\mathbf{u}).$$

If we apply the gradient method to $\bar{f}$, the search direction at a point $\bar{\mathbf{x}}$ (which corresponds to the point $x = P^{-1/2}\bar{\mathbf{x}}$ for the original problem) is

$$\Delta\bar{\mathbf{x}} = -\nabla\bar{f}(\bar{\mathbf{x}}) = -P^{-1/2}\nabla f(P^{-1/2}\bar{\mathbf{x}}) = -P^{-1/2}\nabla f(\mathbf{x}).$$

This gradient search direction corresponds to the direction

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

for the original variable $\mathbf{x}$. So the steepest descent method in the norm $\|\cdot\|_{\mathbf{P}}$ can be regarded as the gradient method applied to the problem after the change of coordinates $\bar{\mathbf{x}} = \mathbf{P}^{1/2}\mathbf{x}$.

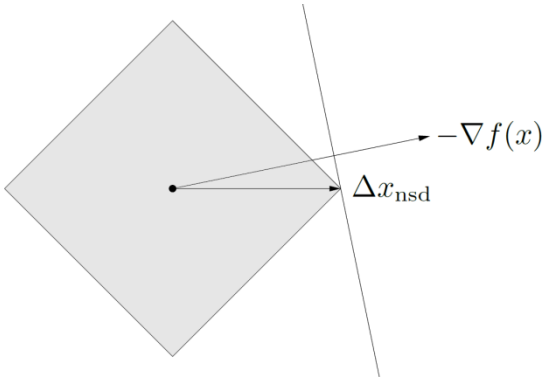
# Descent methods

- Steepest descent for $l_1$-norm

$$\Delta \mathbf{x}_{nsd} = \underset{\mathbf{v}}{\operatorname{argmin}}\{\nabla f(\mathbf{x})^T \mathbf{v} | \|\mathbf{v}\|_1 \leq 1\}$$

can be easily characterized. Let $i$ be any index for which $\|\nabla f(\mathbf{x})\|_\infty = |(\nabla f(\mathbf{x}))_i|$. Then $\Delta \mathbf{x}_{nsd}$ for the $l_1$-norm is given by

$$\Delta \mathbf{x}_{nsd} = -\text{sign}\left(\frac{\partial f(\mathbf{x})}{\partial x_i}\right)\mathbf{e}_i.$$




An unnormalized steepest descent direction is then

$$\Delta \mathbf{x}_{sd} = \Delta \mathbf{x}_{nsd}\|\nabla f(\mathbf{x})\|_\infty = -\frac{\partial f(\mathbf{x})}{\partial x_i}\mathbf{e}_i.$$

not exactly!

Thus, the normalized steepest descent direction in $l_1$-norm can always be chosen along which the approximate decrease in $f$ is greatest.

The algorithm is sometimes called a *coordinate-descent* algorithm. This can greatly simplify, or even trivialize, the line search.

# Descent methods

- Steepest descent for $l_1$-norm

Example: *Frobenius norm scaling.*

$$\min_{\mathbf{d}} \sum_{i,j=1}^{n} M_{ij}^2 d_i^2 / d_j^2.$$

Using the change of variables $x_i = 2 \log d_i$,

$$\min_{\mathbf{x}} f(\mathbf{x}) = \log \left( \sum_{i,j=1}^{n} M_{ij}^2 r^{x_i - x_j} \right).$$

Keeping all components except the $k$th fixed, we can write $f(\mathbf{x}) = \log(\alpha_k + \beta_k e^{x_k} + \gamma_k e^{x_k})$, where

$$\alpha_k = M_{kk}^2 + \sum_{i,j \neq k} M_{ij}^2 e^{x_i - x_j}, \beta_k = \sum_{i \neq k} M_{ij}^2 e^{x_i}, \gamma_k = \sum_{j \neq k} M_{ij}^2 e^{-x_j}.$$

The minimum of $f(\mathbf{x})$, as a function of $x_k$, is obtained for $x_k = \log(\beta_k / \gamma_k)/2$.

# Descent methods

- Steepest descent for $l_1$-norm

The $l_1$-steepest descent algorithm with exact line search consists of repeating the following steps.

1. Compute the gradient $(\nabla f(\mathbf{x}))_i = \frac{\beta_i e^{x_i} + \gamma_i e^{x_i}}{\alpha_i + \beta_i e^{x_i} + \gamma_k e^{x_i}}, i = 1, ..., n.$
2. Select a largest (in absolute value) component of $\nabla f(\mathbf{x})$ : $|\nabla f(\mathbf{x})|_k = \|\nabla f(\mathbf{x})\|_\infty$.
3. Minimize $f$ over the scalar variable $x_k$, by setting $x_k = \log(\beta_k/\gamma_k)/2$.

# Descent methods

- Choice of norm for steepest descent

Prescription for choosing $\mathbf{P}$: It should be chosen so that the sublevel sets of $f$, transformed by $\mathbf{P}^{-1/2}$, are well conditioned. For example if an approximation $\hat{\mathbf{H}}$ of the Hessian at the optimal point $\mathbf{H}(\mathbf{x}^*)$ were known, a very good choice of $\mathbf{P}$ would be $\mathbf{P} = \hat{\mathbf{H}}$, since the Hessian of $\tilde{f}$ at the optimum is then

$$\hat{\mathbf{H}}^{-1/2} \nabla^2 f(\mathbf{x}^*) \hat{\mathbf{H}}^{-1/2} \approx \mathbf{I},$$

and so is likely to have a low condition number. This same idea can be described without a change of coordinates. Saying that a sublevel set has low condition number after the change of coordinates $\hat{\mathbf{x}} = \mathbf{P}^{1/2}\mathbf{x}$ is the same as saying that the ellipsoid

$$\varepsilon = \{\mathbf{x} | \mathbf{x}^T \mathbf{P} \mathbf{x} \leq 1\}$$

approximates the shape of the sublevel set.

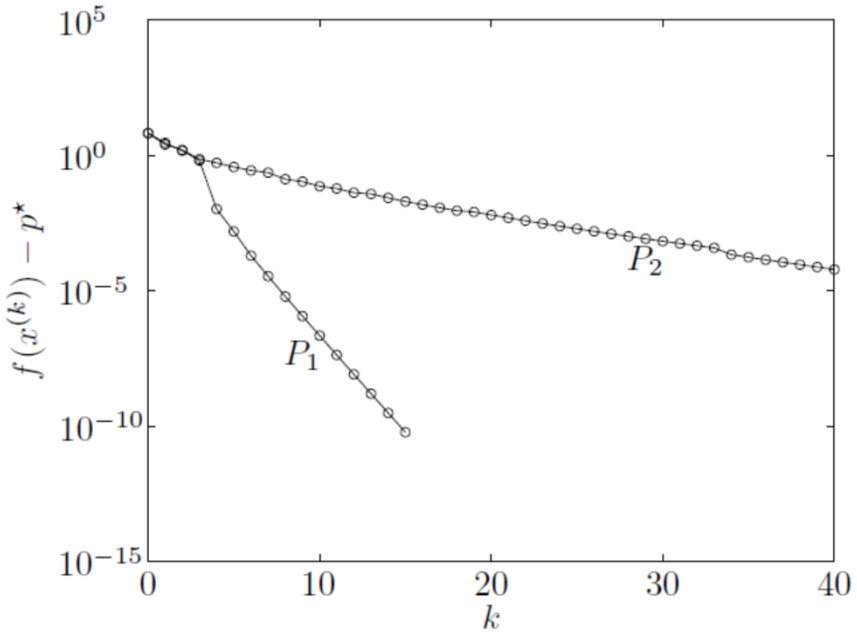
# Descent methods

- Examples

$f(x_1, x_2) = e^{x_1+3x_2-0.1} + e^{x_1-3x_2-0.1} + e^{-x_1-0.1}$.

$$\mathbf{P}_1 = \begin{bmatrix} 2 & 0 \\ 0 & 8 \end{bmatrix}, \mathbf{P}_2 = \begin{bmatrix} 8 & 0 \\ 0 & 2 \end{bmatrix}.$$

In both cases we use a backtracking line search with $\alpha = 0.1$ and $\beta = 0.7$.



# Descent methods

- Examples