A Secant Method for Nonsmooth Optimization

Asef Nazari

CSIRO Melbourne

CARMA Workshop on Optimization, Nonlinear Analysis, Randomness and Risk
Newcastle, Australia
12 July, 2014

The problem

Unconstrained Optimization Problem

$$\min_{x \in R^n} f(x)$$

- $f : R^n \longrightarrow R^n$
- Locally Lipshitz Ö

The problem

Unconstrained Optimization Problem

$$\min_{x \in R^n} f(x)$$

- $f : R^n \longrightarrow R^n$
- Locally Lipshitz Ö
- Why just unconstrained?

Transforming Constrained into Unconstrained

Constrained Optimization Problem

$$\min_{x \in Y} f(x)$$

where $Y \subset R^n$.

Transforming Constrained into Unconstrained

Constrained Optimization Problem

$$\min_{x \in Y} f(x)$$

where $Y \subset R^n$.

- Distance function

$$dist(x, Y) = \min_{y \in Y} \|y - x\|$$

Transforming Constrained into Unconstrained

Constrained Optimization Problem

$$\min_{x \in Y} f(x)$$

where $Y \subset R^n$.

- Distance function

$$dist(x, Y) = \min_{y \in Y} \|y - x\|$$

- Theory of penalty function (under some conditions)

$$\min_{x \in R^n} f(x) + \sigma dist(x, y)$$

- $f(x) + \sigma dist(x, y)$ is a nonsmooth function

Sources of Nonsmooth Problems

- **Minimax Problem**

$$\min_{x \in R^n} f(x)$$

$$f(x) = \max_{1 \leq i \leq m} f_i(x)$$

Sources of Nonsmooth Problems

- **Minimax Problem**

$$\min_{x \in R^n} f(x)$$

$$f(x) = \max_{1 \le i \le m} f_i(x)$$

- **System of Nonlinear Equations**

  $f_i(x) = 0, \quad i = 1, \ldots, m,$

Sources of Nonsmooth Problems

- **Minimax Problem**

$$\min_{x \in R^n} f(x)$$

$$f(x) = \max_{1 \le i \le m} f_i(x)$$

- **System of Nonlinear Equations**

  $f_i(x) = 0, \quad i = 1, \ldots, m,$

  we often do

$$\min_{x \in R^n} \|\bar{f}(x)\|$$

  where $\bar{f}(x) = (f_1(x), \ldots, f_m(x))$

Structure of Optimization Algorithms

Step 1 **Initial Step** $x_0 \in R^n$

Step 2 **Termination Criteria**

Step 3 **Finding descent direction** d_k at x_k

Step 4 **Finding step size** $f(x_k + \alpha_k d_k) < f(x_k)$

Step 5 **Loop** $x_{k+1} = x_k + \alpha_k d_k$ and go to step 2.

Classification of Algorithms Based on d_k and α_k

Directions

- $d_k = -\nabla f(x_k)$ Steepest Descent Method
- $d_k = -H^{-1}(x_k)\nabla f(x_k)$ Newton Method
- $d_k = -B^{-1}\nabla f(x_k)$ Quasi-Newton Method

Classification of Algorithms Based on d_k and α_k

Directions

- $d_k = -\nabla f(x_k)$ Steepest Descent Method
- $d_k = -H^{-1}(x_k)\nabla f(x_k)$ Newton Method
- $d_k = -B^{-1}\nabla f(x_k)$ Quasi-Newton Method

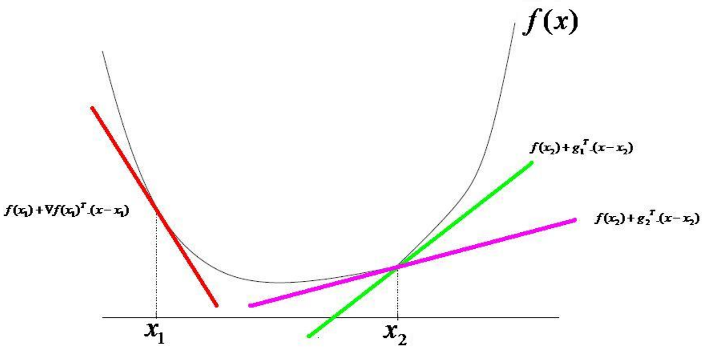
Step sizes

- $h(\alpha) = f(x_k + \alpha d_k)$
  1. exactly solve $h'(\alpha) = 0$ exact line search
  2. loosly solve it, inexact line serach
- Trust region methods

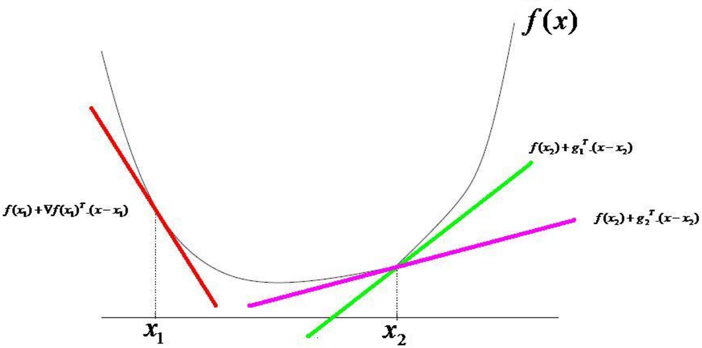
When the objective function is smooth

- Descent direction $f'(x_k, d) < 0$
- $f'(x_k, d) = \langle \nabla f(x_k), d \rangle \leq 0$ descent direction
- $-\nabla f(x_k)$ steepest descent direction
- $\|\nabla f(x_k)\| \leq \varepsilon$ good stopping criteria (First Order Necessary Conditions)

Subgradient



Subgradient



Basic Inequality for convex differentiable function:

$$f(x) \geq f(y) + \nabla f(y)^T.(x-y) \quad \forall x \in \ Dom(f)$$

Subgradient

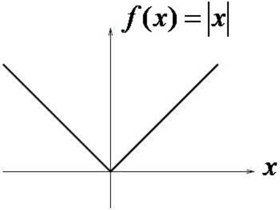
$g \in R^n$ is a subgradient of a convex function f at y if

$$f(x) \geq f(y) + g^T.(x - y) \quad \forall x \in \; Dom(f)$$

Subgradient

$g \in R^n$ is a subgradient of a convex function f at y if

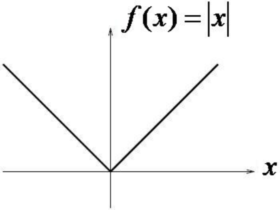
$$f(x) \geq f(y) + g^T.(x - y) \quad \forall x \in \; Dom(f)$$



Subgradient

$g \in R^n$ is a subgradient of a convex function f at y if

$$f(x) \geq f(y) + g^T.(x - y) \quad \forall x \in \ Dom(f)$$

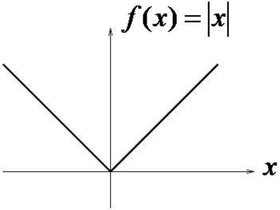


if $x < 0$, $g = -1$

Subgradient

$g \in R^n$ is a subgradient of a convex function f at y if

$$f(x) \geq f(y) + g^T.(x-y) \quad \forall x \in \ Dom(f)$$

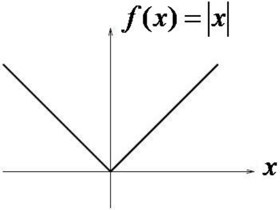


if $x < 0$, $g = -1$
if $x > 0$, $g = 1$

Subgradient

$g \in R^n$ is a subgradient of a convex function f at y if

$$f(x) \geq f(y) + g^T.(x-y) \quad \forall x \in \ Dom(f)$$



if $x < 0$, $g = -1$
if $x > 0$, $g = 1$
if $x = 0$, $g \in [-1, 1]$

Subdifferential

- Set of all subgradients of f at x, $\partial f(x)$, is called subdifferential.

Subdifferential

- Set of all subgradients of f at x, $\partial f(x)$, is called subdifferential.
- for $f(x) = |x|$, the subdifferential at $x = 0$ is $\partial f(0) = [-1, 1]$
  For $x > 0$ $\partial f(x) = \{1\}$
  For $x < 0$ $\partial f(x) = \{-1\}$

Subdifferential

- Set of all subgradients of f at x, $\partial f(x)$, is called subdifferential.
- for $f(x) = |x|$, the subdifferential at $x = 0$ is $\partial f(0) = [-1, 1]$
  For $x > 0$ $\partial f(x) = \{1\}$
  For $x < 0$ $\partial f(x) = \{-1\}$
- f is differentiable at $x \iff \partial f(x) = \{\nabla f(x)\}$

Subdifferential

- Set of all subgradients of f at x, $\partial f(x)$, is called subdifferential.
- for $f(x) = |x|$, the subdifferential at $x = 0$ is $\partial f(0) = [-1, 1]$
  For $x > 0$ $\partial f(x) = \{1\}$
  For $x < 0$ $\partial f(x) = \{-1\}$
- f is differentiable at $x \Longleftrightarrow \partial f(x) = \{\nabla f(x)\}$
- for Lipschitz functions (Clarke)

$$\partial f(x_0) = conv\{\lim \nabla f(x_i) : \quad x_i \longrightarrow x_0 \quad \nabla f(x_i) exists\}$$

Optimality Condition

- For a smooth convex function

$$f(x^*) = \inf_{x \in Dom(f)} f(x) \iff 0 = \nabla f(x^*)$$

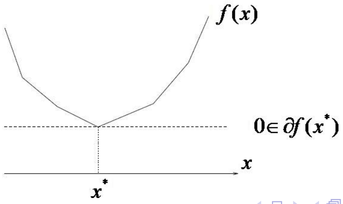
Optimality Condition

- For a smooth convex function

$$f(x^*) = \inf_{x \in Dom(f)} f(x) \Longleftrightarrow 0 = \nabla f(x^*)$$

- For a nonsmooth convex function

$$f(x^*) = \inf_{x \in Dom(f)} f(x) \Longleftrightarrow 0 \in \partial f(x^*)$$



Directional derivative

- In Smooth Case

$$f'(x_k, d) = \lim_{\lambda \to 0^+} \frac{f(x_k + \lambda d) - f(x_k)}{\lambda} = \langle \nabla f(x_k), d \rangle$$

Directional derivative

- In Smooth Case

$$f'(x_k, d) = \lim_{\lambda \to 0^+} \frac{f(x_k + \lambda d) - f(x_k)}{\lambda} = \langle \nabla f(x_k), d \rangle$$

- In nonsmooth case

$$f'(x_k, d) = \lim_{\lambda \to 0^+} \frac{f(x_k + \lambda d) - f(x_k)}{\lambda} = \sup_{g \in \partial f(x_k)} \langle g^T, d \rangle$$

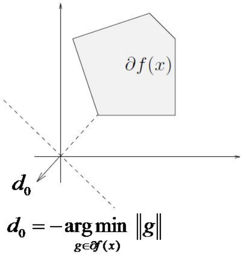
Descent Direction

- In Smooth Case, if $f'(x_k, d) = \langle \nabla f(x_k), d \rangle < 0$. ($-\nabla f(x_k)$ steepest descent direction)

Descent Direction

- In Smooth Case, if $f'(x_k, d) = \langle \nabla f(x_k), d \rangle < 0$. ($-\nabla f(x_k)$ steepest descent direction)
- In nonsmooth case, if $f'(x_k, d) < 0$. It is proved that

$$d = -\underset{g \in \partial f(x_k)}{\operatorname{argmin}} \|g\|$$



In Summary

- Optimality condition

$$f(x^*) = \inf_{x \in Dom(f)} f(x) \Longleftrightarrow 0 \in \partial f(x^*)$$

- Directional Derivative

$$f'(x_k, d) = \lim_{\lambda \to 0^+} \frac{f(x_k + \lambda d) - f(x_k)}{\lambda} = \sup_{g \in \partial f(x_k)} \langle g^T, d \rangle$$

- Steepest Descent Direction

$$d = -\underset{g \in \partial f(x_k)}{argmin} \|g\|$$

Methods for nonsmooth problems

The way we treat $\partial f(x)$ leads to different types of algorithms in nonsmooth optimization

- Subgradient method (one subgradient at each iteration)
- Bundle method (a bundle of subgradients in each iteration)
- Gradient Sampling method
- smoothing technique

Definition of r-secant

- $S_1 = \{g \in R^n : \|g\| = 1\}$ the unit sphere in R^n

Definition of r-secant

- $S_1 = \{g \in R^n : \|g\| = 1\}$ the unit sphere in R^n
- $g \in S_1$ we define

$$g^{max} = \max \{|g_i|, \quad i = 1, \ldots, n\}.$$

Definition of r-secant

- $S_1 = \{g \in R^n : \|g\| = 1\}$ the unit sphere in R^n
- $g \in S_1$ we define

$$g^{max} = \max\{|g_i|, \quad i = 1, \ldots, n\}.$$

- $g \in S_1$ and $g_j = g^{max}$, $v \in \partial f(x + rg)$

Definition of r-secant

- $S_1 = \{g \in R^n : \|g\| = 1\}$ the unit sphere in R^n
- $g \in S_1$ we define

$$g^{max} = \max\{|g_i|, \quad i = 1, \dots, n\}.$$

- $g \in S_1$ and $g_j = g^{max}$, $v \in \partial f(x + rg)$
- $s = s(x, g, r) \in R^n$ where

$$s = (s_1, \dots, s_n) : s_i = v_i, \ i = 1, \dots, n, \ i \neq j$$

and

$$s_j = \frac{f(x + rg) - f(x) - r\sum_{i=1, i\neq j}^{n} s_i g_i}{rg_j}$$

is called an r-secant of the function f at a point x in the direction g.

Facts about r-secant

- If $n = 1$,

$$s = \frac{f(x + rg) - f(x)}{rg}$$

Facts about r-secant

- If $n = 1$,

$$s = \frac{f(x + rg) - f(x)}{rg}$$

- Mean Value Theorem for r-secants

$$f(x + rg) - f(x) = r\langle s(x, g, r), g\rangle$$

Facts about r-secant

- If $n = 1$,

$$s = \frac{f(x + rg) - f(x)}{rg}$$

- Mean Value Theorem for r-secants

$$f(x + rg) - f(x) = r\langle s(x, g, r), g\rangle$$

- Set of all possible r-secants of the function f at the point x

$$S_r f(x) = \{s \in R^n : \exists g \in S_1 : \quad s = s(x, g, r)\}$$

Facts about r-secant

- If $n = 1$,

$$s = \frac{f(x+rg) - f(x)}{rg}$$

- Mean Value Theorem for r-secants

$$f(x+rg) - f(x) = r\langle s(x,g,r), g\rangle$$

- Set of all possible r-secants of the function f at the point x

$$S_r f(x) = \{s \in R^n : \exists g \in S_1 : \quad s = s(x,g,r)\}$$

1. compact for $r > 0$

Facts about r-secant

- If $n = 1$,

$$s = \frac{f(x + rg) - f(x)}{rg}$$

- Mean Value Theorem for r-secants

$$f(x + rg) - f(x) = r\langle s(x, g, r), g\rangle$$

- Set of all possible r-secants of the function f at the point x

$$S_r f(x) = \{s \in R^n : \exists g \in S_1 : \quad s = s(x, g, r)\}$$

1. compact for $r > 0$
2. $(x, r) \mapsto S_r f(x), r > 0$ is closed and upper semi-continuous

A new set

- $S_0^c f(x) = conv\{v \in R^n : \exists (g \in S_1,\ r_k \to +0,\ k \to +\infty) :$

$$v = \lim_{k \to +\infty} s(x, g, r_k)\},$$

A new set

- $S_0^c f(x) = conv\{v \in R^n : \exists (g \in S_1,\ r_k \to +0,\ k \to +\infty) :$

$$v = \lim_{k \to +\infty} s(x, g, r_k)\},$$

- For regular and semismooth function f at a point $x \in R^n$:

$$\partial f(x) = S_0^c f(x).$$

Optimality condition

- Let $x \in R^n$ be a local minimizer of the function f and it is directionally differentiable at x. Then

$$0 \in S_0^c f(x).$$

- $x \in R^n$ is an r-stationary point for a function f on R^n if $0 \in S_r^c f(x)$.
- $x \in R^n$ is an (r, δ)-stationary point for a function f on R^n if $0 \in S_r^c f(x) + B_\delta$ where

$$B_\delta = \{v \in R^n : \quad \|v\| \leq \delta\}.$$

Descent Direction

- If $x \in R^n$ is not an r-stationary point of a function f on R^n,

$$0 \notin S_r^c f(x).$$

we can compute a descent direction using the set $S_r^c f(x)$

Descent Direction

- If $x \in R^n$ is not an r-stationary point of a function f on R^n,

$$0 \notin S_r^c f(x).$$

  we can compute a descent direction using the set $S_r^c f(x)$
- Let $x \in R^n$ and for given $r > 0$

$$\min\{\|v\| : v \in S_r^c f(x)\} = \|v^0\| > 0.$$

  Then for $g^0 = -\|v^0\|^{-1} v^0$

$$f(x + rg^0) - f(x) \leq -r\|v^0\|.$$

Descent Direction

- If $x \in R^n$ is not an r-stationary point of a function f on R^n,

$$0 \notin S_r^c f(x).$$

  we can compute a descent direction using the set $S_r^c f(x)$
- Let $x \in R^n$ and for given $r > 0$

$$\min\{\|v\| : v \in S_r^c f(x)\} = \|v^0\| > 0.$$

  Then for $g^0 = -\|v^0\|^{-1} v^0$

$$f(x + rg^0) - f(x) \le -r\|v^0\|.$$

-

$$\begin{aligned} &minimize \qquad \|v\|^2 \\ &subjectto \quad v \in S_r^c f(x). \end{aligned}$$

An algorithm for descent direction (Alg1)

step 1. compute an r-secant $s^1 = s(x, g^1, r)$. Set $\overline{W}_1(x) = \{s^1\}$ and $k = 1$.

An algorithm for descent direction (Alg1)

step 1. compute an r-secant $s^1 = s(x, g^1, r)$. Set $\overline{W}_1(x) = \{s^1\}$ and $k = 1$.

step 2. $\|w^k\|^2 = \min\{\|w\|^2 : w \in co\overline{W}_k(x)\}$. If

$$\|w^k\| \leq \delta,$$

then stop. Otherwise go to Step 3.

An algorithm for descent direction (Alg1)

step 1. compute an r-secant $s^1 = s(x, g^1, r)$. Set $\overline{W}_1(x) = \{s^1\}$ and $k = 1$.

step 2. $\|w^k\|^2 = \min\{\|w\|^2 : w \in co\overline{W}_k(x)\}$. If

$$\|w^k\| \leq \delta,$$

then stop. Otherwise go to Step 3.

step 3. $g^{k+1} = -\|w^k\|^{-1} w^k$.

An algorithm for descent direction (Alg1)

step 1. compute an r-secant $s^1 = s(x, g^1, r)$. Set $\overline{W}_1(x) = \{s^1\}$ and $k = 1$.

step 2. $\|w^k\|^2 = \min\{\|w\|^2 : w \in co\overline{W}_k(x)\}$. If

$$\|w^k\| \leq \delta,$$

then stop. Otherwise go to Step 3.

step 3. $g^{k+1} = -\|w^k\|^{-1} w^k$.

step 4. $f(x + rg^{k+1}) - f(x) \leq -cr\|w^k\|$, then stop. Otherwise go to Step 5.

An algorithm for descent direction (Alg1)

step 1. compute an r-secant $s^1 = s(x, g^1, r)$. Set $\overline{W}_1(x) = \{s^1\}$ and $k = 1$.

step 2. $\|w^k\|^2 = \min\{\|w\|^2 : w \in co\overline{W}_k(x)\}$. If

$$\|w^k\| \leq \delta,$$

then stop. Otherwise go to Step 3.

step 3. $g^{k+1} = -\|w^k\|^{-1}w^k$.

step 4. $f(x + rg^{k+1}) - f(x) \leq -cr\|w^k\|$, then stop. Otherwise go to Step 5.

step 5. $s^{k+1} = s(x, g^{k+1}, r)$ with respect to the direction g^{k+1}, construct the set $\overline{W}_{k+1}(x) = co\left\{\overline{W}_k(x) \bigcup \{s^{k+1}\}\right\}$, set $k = k + 1$ and go to Step 2.

The secant method (r, δ)-stationary point(Alg 2)

step 1. $x^0 \in R^n$ and set $k = 0$.

The secant method (r, δ)-stationary point(Alg 2)

step 1. $x^0 \in R^n$ and set $k = 0$.

step 2. Apply Alg 1 at $x = x^k$
Either $\|v^k\| \leq \delta$ or for the search direction $g^k = -\|v^k\|^{-1}v^k$

The secant method (r, δ)-stationary point(Alg 2)

step 1. $x^0 \in R^n$ and set $k = 0$.

step 2. Apply Alg 1 at $x = x^k$
Either $\|v^k\| \leq \delta$ or for the search direction $g^k = -\|v^k\|^{-1} v^k$

step 3. If $\|v^k\| \leq \delta$ then stop. Otherwise go to Step 4.

The secant method (r, δ)-stationary point(Alg 2)

step 1. $x^0 \in R^n$ and set $k = 0$.

step 2. Apply Alg 1 at $x = x^k$
Either $\|v^k\| \leq \delta$ or for the search direction $g^k = -\|v^k\|^{-1} v^k$

step 3. If $\|v^k\| \leq \delta$ then stop. Otherwise go to Step 4.

step 4. $x^{k+1} = x^k + \sigma_k g^k$, where σ_k is defined as follows

$$\sigma_k = arg \max \left\{ \sigma \geq 0 : \; f(x^k + \sigma g^k) - f(x^k) \leq -c_2 \sigma \|v^k\| \right\}.$$

Set $k = k + 1$ and go to Step 2.

The secant method (Alg 3)

step 1. $x^0 \in R^n$ and set $k = 0$.

The secant method (Alg 3)

step 1. $x^0 \in R^n$ and set $k = 0$.

step 2. Apply Alg 2 to x^k for $r = r_k$ and $\delta = \delta_k$. This algorithm terminates after a finite number of iterations $p > 0$ and as a result the algorithm finds (r_k, δ_k)-stationary point x^{k+1}.

The secant method (Alg 3)

step 1. $x^0 \in R^n$ and set $k = 0$.

step 2. Apply Alg 2 to x^k for $r = r_k$ and $\delta = \delta_k$. This algorithm terminates after a finite number of iterations $p > 0$ and as a result the algorithm finds (r_k, δ_k)-stationary point x^{k+1}.

step 3. Set $k = k + 1$ and go to Step 2.

Convergence Theorem

Theorem

Assume that the function f is locally Lipschitz and the set $\mathcal{L}(x^0)$ is bounded for starting points $x^0 \in R^n$. Then every accumulation point of the sequence $\{x^k\}$ belongs to the set
$X^0 = \{x \in R^n : 0 \in \partial f(x)\}$.

Results

Prob.	Secant			Bundle		
	f_{av}	n_b	n_s	f_{av}	n_b	n_s
P1	1.95222	20	20	1.95222	20	20
P2	-44	20	20	-44	20	20
P3	3.70348	20	20	3.70348	20	20
P4	0.90750	17	17	0.90750	17	17
P5	0.57592	4	4	0.00000	20	20
P6	3.59972	20	20	3.59972	20	20
P7	-44	20	20	-28.14011	12	12
P8	0.03565	7	13	0.03051	9	17
P9	0.02886	0	7	0.01520	2	16
P10	115.70644	20	20	115.70644	20	20
P11	0.00291	0	0	0.00264	20	20
P12	0.01773	0	6	0.02752	14	16
P13	0.10916	3	9	0.30582	3	15
P14	0.24037	8	18	0.32527	5	9
P15	0.03490	20	20	0.30572	12	12
P16	0.12402	0	11	0.39131	2	9
P17	680.63006	20	20	680.63006	20	20
P18	24.30621	20	20	24.30621	20	20
P19	93.90566	20	20	93.90525	20	20
P20	0.00302	0	0	0.00000	20	20
P21	0.21456	0	9	0.22057	1	14
P22	2.00000	20	20	2.00000	20	20
P23	0.10000	18	18	0.07607	17	17
P24	1.50000	7	18	2.30008	2	10
P25	0.00000	20	20	0.00000	20	20
P26	0.00000	20	20	0.00000	20	20
P27	24.72876	0	18	35.19230	0	2
P28	8.53416×10^6	10	18	10.29861×10^6	5	12
P29	5.56355×10^6	3	20	7.10874×10^6	0	9
P30	2.93562×10^6	2	19	3.16944×10^6	0	2
P31	7.27436×10^5	11	19	7.26213×10^5	12	20
P32	3.94879×10^5	10	19	3.94922×10^5	5	13
P33	1.86467×10^5	13	18	1.88295×10^5	6	9

Results

Prob.	Secant			Bundle		
	n_f	n_{sub}	t	n_f	n_{sub}	t
P1	221	160	0.000	10	10	0.001
P2	1113	601	0.002	12	11	0.001
P3	974	701	0.143	65	55	0.003
P4	749	593	0.002	22	9	0.000
P5	4735	461	0.002	493	251	0.001
P6	574	309	0.001	20	16	0.000
P7	1225	579	0.003	146	56	0.001
P8	1184	342	0.009	1626	171	0.009
P9	2968	601	0.003	57891	4843	0.186
P10	1192	619	0.010	29	15	0.001
P11	1829	811	0.005	26081	1904	0.122
P12	4668	1327	0.014	372	173	0.007
P13	1310	749	0.014	61	26	0.002
P14	837	686	0.020	51	23	0.002
P15	1373	1090	0.085	4985	445	0.108
P16	3463	1989	0.302	60331	4761	1.099
P17	1270	860	0.012	58	33	0.000
P18	2234	1592	0.029	18	15	0.000
P19	4752	4001	0.362	35	26	0.003
P20	3399	2733	3.393	160	52	0.030
P21	1529	849	0.249	66280	4974	2.971
P22	289	198	0.001	32	32	0.000
P23	2278	277	0.001	22	22	0.000
P24	470	450	0.003	37	37	0.000
P25	1022	983	0.010	25	25	0.001
P26	2677	920	0.005	68	68	0.001
P27	725	707	0.039	37	37	0.002
P28	305	289	0.099	16	16	0.007
P29	417	401	0.360	21	21	0.022
P30	766	736	2.734	51	51	0.205
P31	283	257	0.254	19	19	0.024
P32	406	370	0.955	28	28	0.085
P33	703	653	7.003	67	67	0.763