Quickselect con muestreo adaptativo

C. Martínez D. Panario A. Viola

Univ. Politècnica de Catalunya
Carleton University (Canadá)
Univ. de la República (Uruguay)

IV Jornadas de Matemática Discreta y Algorítmica
Septiembre 2004

Introducción

- Quickselect (Hoare, 1962) selecciona el m-ésimo menor elemento de entre n elementos
- Particiona el vector dado en torno a un pivote y continúa la búsqueda, recursivamente, en el subvector apropiado
- Quickselect es eficiente (Knuth, 1971)

$$C_{n,m} = m_0(\alpha) \cdot n + o(n) = 2(1 + \mathcal{H}(\alpha)) \cdot n + o(n)$$
$$= (2 - 2(\alpha \ln \alpha + (1 - \alpha) \ln(1 - \alpha))) \cdot n + o(n),$$

donde $0 \leq \alpha = \frac{m}{n} \leq 1$

Introducción

- Quickselect (Hoare, 1962) selecciona el m-ésimo menor elemento de entre n elementos
- Particiona el vector dado en torno a un pivote y continúa la búsqueda, recursivamente, en el subvector apropiado
- Quickselect es eficiente (Knuth, 1971)

$$C_{n,m} = m_0(\alpha) \cdot n + o(n) = 2(1 + \mathcal{H}(\alpha)) \cdot n + o(n)$$
$$= (2 - 2(\alpha \ln \alpha + (1 - \alpha) \ln(1 - \alpha))) \cdot n + o(n),$$

donde $0 \leq \alpha = \frac{m}{n} \leq 1$

Introducción

- Quickselect (Hoare, 1962) selecciona el m-ésimo menor elemento de entre n elementos
- Particiona el vector dado en torno a un pivote y continúa la búsqueda, recursivamente, en el subvector apropiado
- Quickselect es eficiente (Knuth, 1971)

$$\begin{aligned} C_{n,m} &= m_0(\alpha) \cdot n + o(n) = 2(1 + \mathcal{H}(\alpha)) \cdot n + o(n) \\ &= (2 - 2(\alpha \ln \alpha + (1-\alpha)\ln(1-\alpha))) \cdot n + o(n), \end{aligned}$$

donde $0 \leq \alpha = \frac{m}{n} \leq 1$

El algoritmo

```
Elem quickselect(vector<Elem>& A, int m) {
    int l = 0; int u = A.size() - 1;
    int k, p;
    while (l <= u) {
       p = selecciona_pivote(A, l, u, m);
       swap(A[p], A[l]);
       particiona(A, l, u, k);
       if (m < k) u = k-1;
       else if (m > k) l = k+1;
       else return A[k];
}   }
```

Mediana-de-$(2t+1)$

- Usar una muestra de talla $s = 2t + 1$ para seleccionar el pivote de cada iteración mejora el rendimiento promedio y reduce la probabilidad del caso peor
- Para quickselect con mediana-de-3 (Kirschenhofer, Martínez, Prodinger, 1995)

$$m_1(\alpha) = 2 + 3\alpha(1 - \alpha)$$

- Para toda α, $0 \leq \alpha \leq 1$, $m_0(\alpha) \geq m_1(\alpha)$. Además $\overline{m}_0 = 3$ y $\overline{m}_1 = 2{,}5$

Mediana-de-$(2t+1)$

- Usar una muestra de talla $s = 2t + 1$ para seleccionar el pivote de cada iteración mejora el rendimiento promedio y reduce la probabilidad del caso peor
- Para quickselect con mediana-de-3 (Kirschenhofer, Martínez, Prodinger, 1995)

$$m_1(\alpha) = 2 + 3\alpha(1 - \alpha)$$

- Para toda α, $0 \leq \alpha \leq 1$, $m_0(\alpha) \geq m_1(\alpha)$. Además $\overline{m}_0 = 3$ y $\overline{m}_1 = 2{,}5$

Mediana-de-$(2t+1)$

- Usar una muestra de talla $s = 2t + 1$ para seleccionar el pivote de cada iteración mejora el rendimiento promedio y reduce la probabilidad del caso peor
- Para quickselect con mediana-de-3 (Kirschenhofer, Martínez, Prodinger, 1995)

$$m_1(\alpha) = 2 + 3\alpha(1 - \alpha)$$

- Para toda α, $0 \leq \alpha \leq 1$, $m_0(\alpha) \geq m_1(\alpha)$. Además $\overline{m}_0 = 3$ y $\overline{m}_1 = 2{,}5$

Muestreo adaptativo

- Usar la mediana de muestras no es lo más natural
- Por ejemplo, si buscamos el 110-ésimo de entre 10000 es más lógico tomar el menor de una muestra de tres!
- En proporcional-de-3 se usan muestras de tres elementos. Si el rango relativo es $\leq 1/3$ se toma el menor de la muestra, si está entre $1/3$ y $2/3$ se toma la mediana, y si es $\geq 2/3$ se toma el mayor de la muestra.

Muestreo adaptativo

- Usar la mediana de muestras no es lo más natural
- Por ejemplo, si buscamos el 110-ésimo de entre 10000 es más lógico tomar el menor de una muestra de tres!
- En proporcional-de-3 se usan muestras de tres elementos. Si el rango relativo es $\leq 1/3$ se toma el menor de la muestra, si está entre $1/3$ y $2/3$ se toma la mediana, y si es $\geq 2/3$ se toma el mayor de la muestra.

Muestreo adaptativo

- Usar la mediana de muestras no es lo más natural
- Por ejemplo, si buscamos el 110-ésimo de entre 10000 es más lógico tomar el menor de una muestra de tres!
- En proporcional-de-3 se usan muestras de tres elementos. Si el rango relativo es $\leq 1/3$ se toma el menor de la muestra, si está entre $1/3$ y $2/3$ se toma la mediana, y si es $\geq 2/3$ se toma el mayor de la muestra.

Muestreo adaptativo

Ejemplo: Buscamos el cuarto ($m = 4$) de entre $n = 15$ elementos

9	5	10	12	3	1	11	15	7	2	8	13	6	4	14

$$\alpha = 4/15 < 1/3$$

Muestreo adaptativo

Ejemplo: Buscamos el cuarto ($m = 4$) de entre $n = 15$ elementos

9	5	10	12	3	1	11	15	7	2	8	13	6	4	14

$$\alpha = 4/15 < 1/3$$

Muestreo adaptativo

Ejemplo: Buscamos el cuarto ($m = 4$) de entre $n = 15$ elementos

9	5	10	12	3	1	11	15	7	2	8	13	6	4	14

$$\alpha = 4/15 < 1/3$$

Muestreo adaptativo

Ejemplo: Buscamos el cuarto ($m = 4$) de entre $n = 15$ elementos

7	5	4	6	3	1	8	2	9	15	11	13	12	10	14

Muestreo adaptativo

Ejemplo: Buscamos el cuarto ($m = 4$) de entre $n = 15$ elementos

7	5	4	6	3	1	8	2	9	15	11	13	12	10	14

$$1/3 < \alpha = 4/8 = 1/2 < 2/3$$

Muestreo adaptativo

Ejemplo: Buscamos el cuarto ($m = 4$) de entre $n = 15$ elementos

7	5	4	6	3	1	8	2	9	15	11	13	12	10	14

$$1/3 < \alpha = 4/8 = 1/2 < 2/3$$

Muestreo adaptativo

Ejemplo: Buscamos el cuarto ($m = 4$) de entre $n = 15$ elementos

1	5	4	2	3	6	8	7	9	15	11	13	12	10	14

Muestreo adaptativo

Ejemplo: Buscamos el cuarto ($m = 4$) de entre $n = 15$ elementos

1	5	4	2	3	6	8	7	9	15	11	13	12	10	14

$$\alpha = 4/5 > 2/3$$

Muestreo adaptativo

Ejemplo: Buscamos el cuarto ($m = 4$) de entre $n = 15$ elementos

1	5	4	2	3	6	8	7	9	15	11	13	12	10	14

$$\alpha = 4/5 > 2/3$$

Muestreo adaptativo

Ejemplo: Buscamos el cuarto ($m = 4$) de entre $n = 15$ elementos

2	3	1	4	5	6	8	7	9	15	11	13	12	10	14

Muestreo adaptativo

- Usar el elemento de la muestra con rango relativo próximo a $\alpha = m/n$
- En general: $r(\alpha) =$ rango del pivote con respecto a la muestra, cuando seleccionamos el m-ésimo de entre n y $\alpha = m/n$
- Dividir $[0, 1]$ en ℓ intervalos con extremos

$$0 = a_0 < a_1 < a_2 < \cdots < a_\ell = 1$$

Sea r_k el valor de $r(\alpha)$ cuando α pertenece al k-ésimo intervalo

Muestreo adaptativo

- Usar el elemento de la muestra con rango relativo próximo a $\alpha = m/n$
- En general: $r(\alpha) =$ rango del pivote con respecto a la muestra, cuando seleccionamos el m-ésimo de entre n y $\alpha = m/n$
- Dividir $[0, 1]$ en ℓ intervalos con extremos

$$0 = a_0 < a_1 < a_2 < \cdots < a_\ell = 1$$

Sea r_k el valor de $r(\alpha)$ cuando α pertenece al k-ésimo intervalo

Muestreo adaptativo

- Usar el elemento de la muestra con rango relativo próximo a $\alpha = m/n$
- En general: $r(\alpha) =$ rango del pivote con respecto a la muestra, cuando seleccionamos el m-ésimo de entre n y $\alpha = m/n$
- Dividir $[0, 1]$ en ℓ intervalos con extremos

$$0 = a_0 < a_1 < a_2 < \cdots < a_\ell = 1$$

Sea r_k el valor de $r(\alpha)$ cuando α pertenece al k-ésimo intervalo

Muestreo adaptativo

- Para mediana-de-$(2t+1)$: $\ell = 1$ y $r_1 = t+1$
- Para proporcional-de-s: $\ell = s$, $a_k = k/s$ y $r_k = k$
- Estrategias del tipo "proporcional-de": $\ell = s$ y $r_k = k$, pero los extremos de los intervalos $a_k \neq k/s$
- Una estrategia de muestreo es simétrica si

$$r(\alpha) = s + 1 - r(1-\alpha)$$

Muestreo adaptativo

- Para mediana-de-$(2t+1)$: $\ell = 1$ y $r_1 = t+1$
- Para proporcional-de-s: $\ell = s$, $a_k = k/s$ y $r_k = k$
- Estrategias del tipo "proporcional-de": $\ell = s$ y $r_k = k$, pero los extremos de los intervalos $a_k \neq k/s$
- Una estrategia de muestreo es simétrica si

$$r(\alpha) = s + 1 - r(1-\alpha)$$

Muestreo adaptativo

- Para mediana-de-$(2t+1)$: $\ell = 1$ y $r_1 = t+1$
- Para proporcional-de-s: $\ell = s$, $a_k = k/s$ y $r_k = k$
- Estrategias del tipo "proporcional-de": $\ell = s$ y $r_k = k$, pero los extremos de los intervalos $a_k \neq k/s$
- Una estrategia de muestreo es simétrica si

$$r(\alpha) = s + 1 - r(1-\alpha)$$

Muestreo adaptativo

- Para mediana-de-$(2t+1)$: $\ell = 1$ y $r_1 = t+1$
- Para proporcional-de-s: $\ell = s$, $a_k = k/s$ y $r_k = k$
- Estrategias del tipo "proporcional-de": $\ell = s$ y $r_k = k$, pero los extremos de los intervalos $a_k \neq k/s$
- Una estrategia de muestreo es simétrica si

$$r(\alpha) = s + 1 - r(1-\alpha)$$

La recurrencia

- Probabilidad de que el r-ésimo elemento de una muestra de tamaño s sea el j-ésimo elemento de los n elementos dados:

$$\pi_{n,j}^{(s,r)} = \frac{\binom{j-1}{r-1}\binom{n-j}{s-r}}{\binom{n}{s}}, \quad 1 \leq r \leq s \leq n, \quad 1 \leq j \leq n$$

La recurrencia

- Número medio de comparaciones $C_{n,m}$ para seleccionar el m-ésimo de entre n:

$$C_{n,m} = n + \Theta(1) + \sum_{j=m+1}^{n} \pi_{n,j}^{(s,r)} \cdot C_{j-1,m} + \sum_{j=1}^{m-1} \pi_{n,j}^{(s,r)} \cdot C_{n-j,m-j}$$

Un teorema general

Teorema

Sea $f(\alpha) = \lim_{n\to\infty, m/n\to\alpha} \frac{C_{n,m}}{n}$. *Entonces*

$$f(\alpha) = 1 + \frac{s!}{(r(\alpha)-1)!(s-r(\alpha))!} \times \left[\int_{\alpha}^{1} f\left(\frac{\alpha}{x}\right) x^{r(\alpha)} (1-x)^{s-r(\alpha)}\, dx + \int_{0}^{\alpha} f\left(\frac{\alpha - x}{1-x}\right) x^{r(\alpha)-1} (1-x)^{s+1-r(\alpha)}\, dx \right].$$

Dos proposiciones elementales

- Si $r(\alpha)$ es simétrica entonces $f(\alpha) = f(1-\alpha)$
- Sea $r_0 = \lim_{\alpha \to 0} r(\alpha)$. Entonces

$$\lim_{\alpha \to 0} f(\alpha) = \frac{s+1}{s+1-r_0}$$

Para las estrategias del tipo "proporcional-de-s" tenemos $r_0 = 1$ y por tanto, $f(0) = 1 + 1/s$; para mediana-de-$(2t+1)$ tenemos $m_t(0) = 2$

Dos proposiciones elementales

- Si $r(\alpha)$ es simétrica entonces $f(\alpha) = f(1-\alpha)$
- Sea $r_0 = \lim_{\alpha \to 0} r(\alpha)$. Entonces

$$\lim_{\alpha \to 0} f(\alpha) = \frac{s+1}{s+1-r_0}$$

Para las estrategias del tipo "proporcional-de-s" tenemos $r_0 = 1$ y por tanto, $f(0) = 1 + 1/s$; para mediana-de-$(2t+1)$ tenemos $m_t(0) = 2$

La ecuación diferencial general

Lema

Sea f_k la restricción de $f(\alpha)$ en el k-ésimo intervalo. Para cualquier estrategia de muestreo adaptativo

$$\frac{d^{s+2}}{d\alpha^{s+2}} f_k(\alpha) = \frac{(-1)^{s+1-r_k} \cdot s!}{\alpha^{s+1-r_k}(r_k-1)!} \frac{d^{r_k+1}}{d\alpha^{r_k+1}} f_k(\alpha) + \frac{s!}{(1-\alpha)^{r_k}(s-r_k)!} \frac{d^{s+2-r_k}}{d\alpha^{s+2-r_k}} f_k(\alpha).$$

Proporcional-de-2

- La ecuación diferencial es

$$\frac{d^2\phi_1}{dx^2} - \frac{2}{1-x}\frac{d\phi_1}{dx} - \frac{2}{x^2}\phi_1 = 0$$

donde $\phi_1(x) = f_1''(x)$ y $f_2(x) = f_1(1-x)$

- La solución es

$$f_1(x) = a\left((x-1)\ln(1-x) + \frac{x^3}{6} + \frac{x^2}{2} - x\right) - b(1+\mathcal{H}(x)) + cx + d$$

Proporcional-de-2

- La ecuación diferencial es

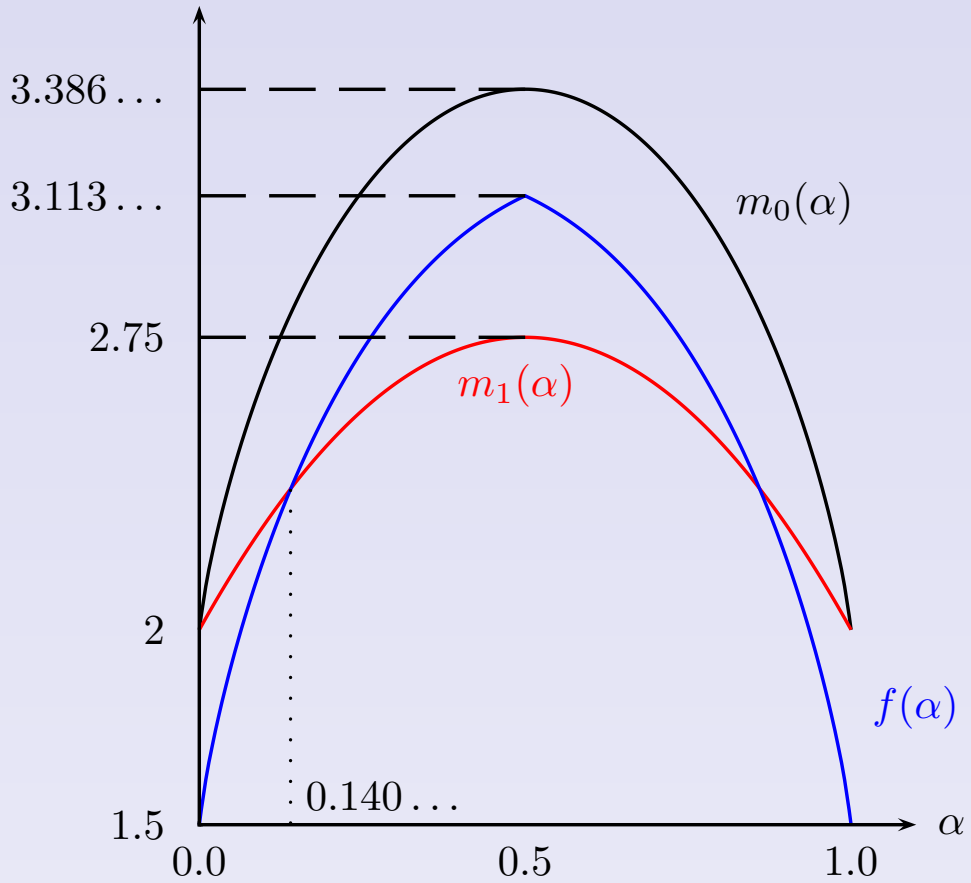
$$\frac{d^2\phi_1}{dx^2} - \frac{2}{1-x}\frac{d\phi_1}{dx} - \frac{2}{x^2}\phi_1 = 0$$

donde $\phi_1(x) = f_1''(x)$ y $f_2(x) = f_1(1-x)$

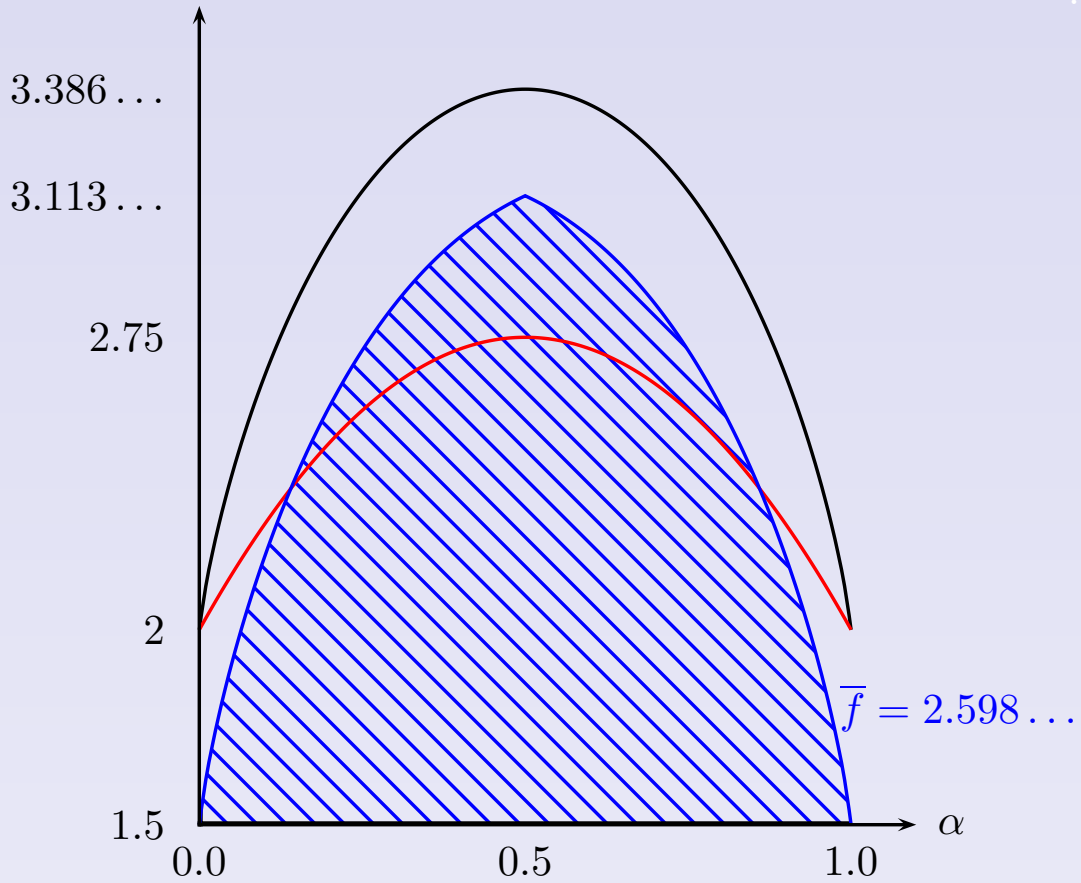
- La solución es

$$f_1(x) = a\left((x-1)\ln(1-x) + \frac{x^3}{6} + \frac{x^2}{2} - x\right) - b(1+\mathcal{H}(x)) + cx + d$$

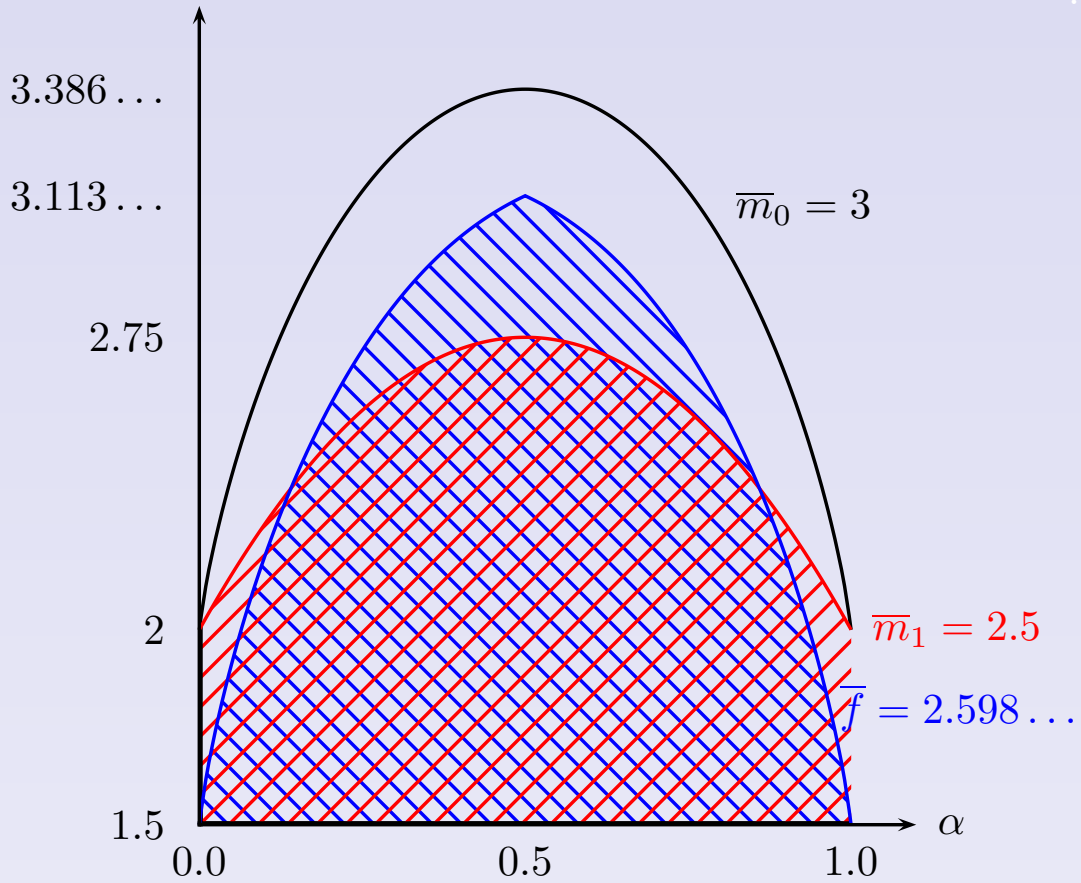
Proporcional-de-2



Proporcional-de-2



Proporcional-de-2



Proporcional-de-3

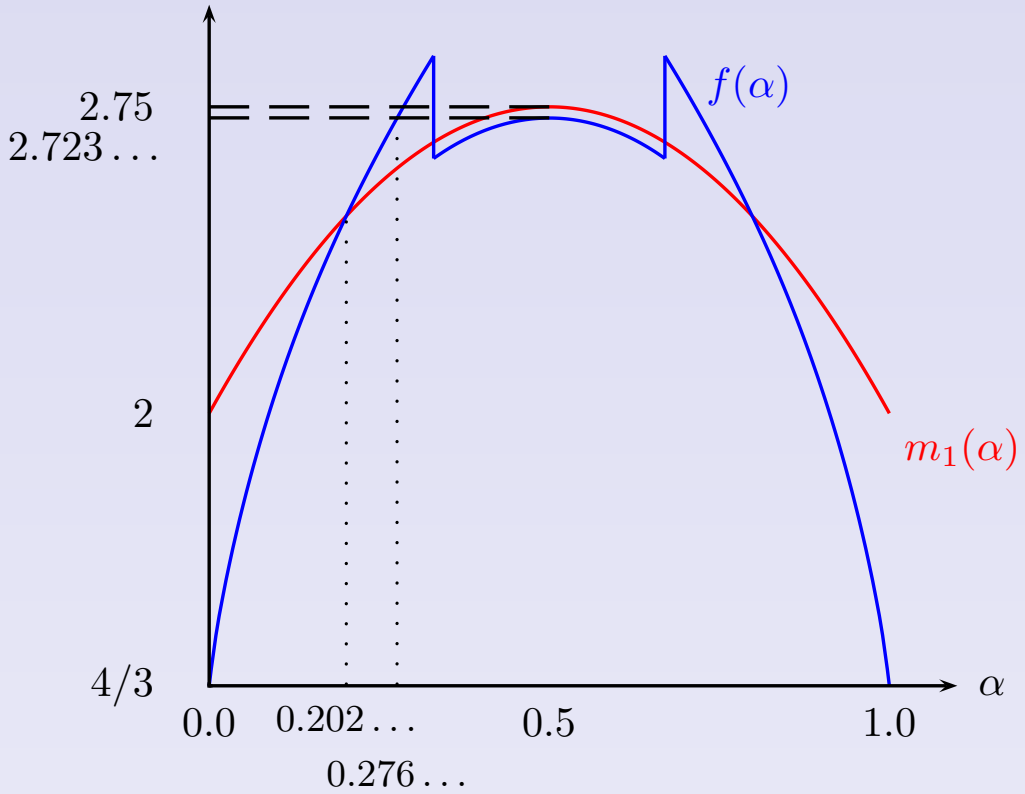
- Para proporcional-de-3,

$$f_1(x) = -C_0(1 + \mathcal{H}(x)) + C_1 + C_2 x + C_3 K_1(x) + C_4 K_2(x),$$
$$f_2(x) = -C_5(1 + \mathcal{H}(x)) + C_6 x(1 - x) + C_7,$$

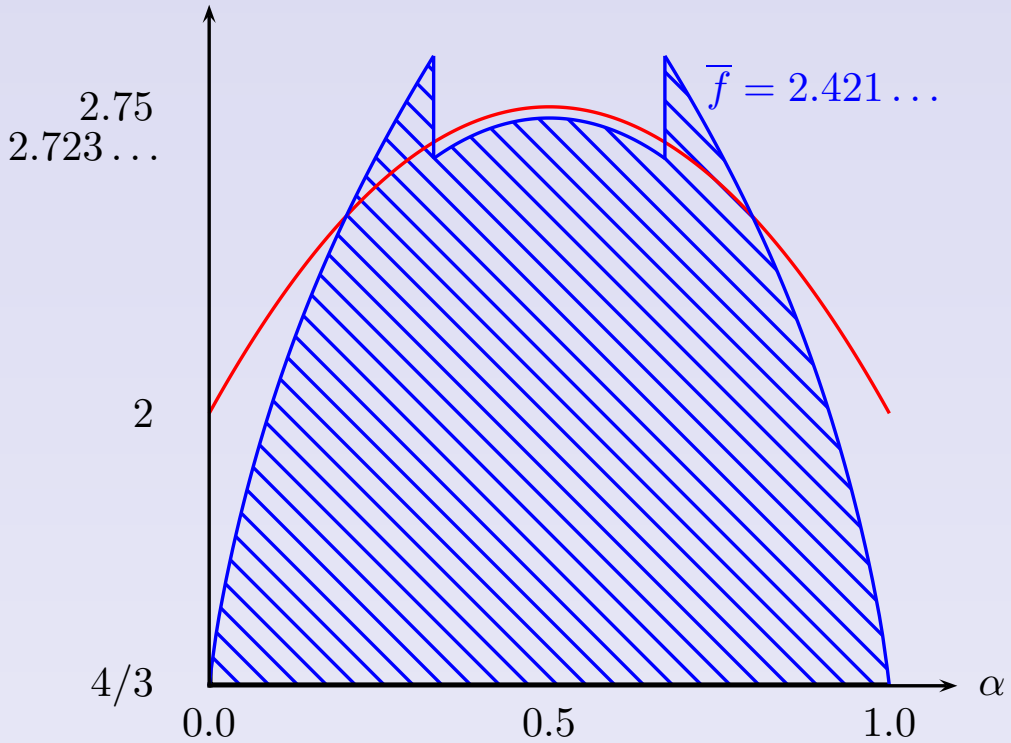
donde

$$K_1(x) = \cos(\sqrt{2}\ln x) \cdot \sum_{n \geq 0} A_n x^{n+4} + \sin(\sqrt{2}\ln x) \cdot \sum_{n \geq 0} B_n x^{n+4},$$
$$K_2(x) = \sin(\sqrt{2}\ln x) \cdot \sum_{n \geq 0} A_n x^{n+4} - \cos(\sqrt{2}\ln x) \cdot \sum_{n \geq 0} B_n x^{n+4}$$

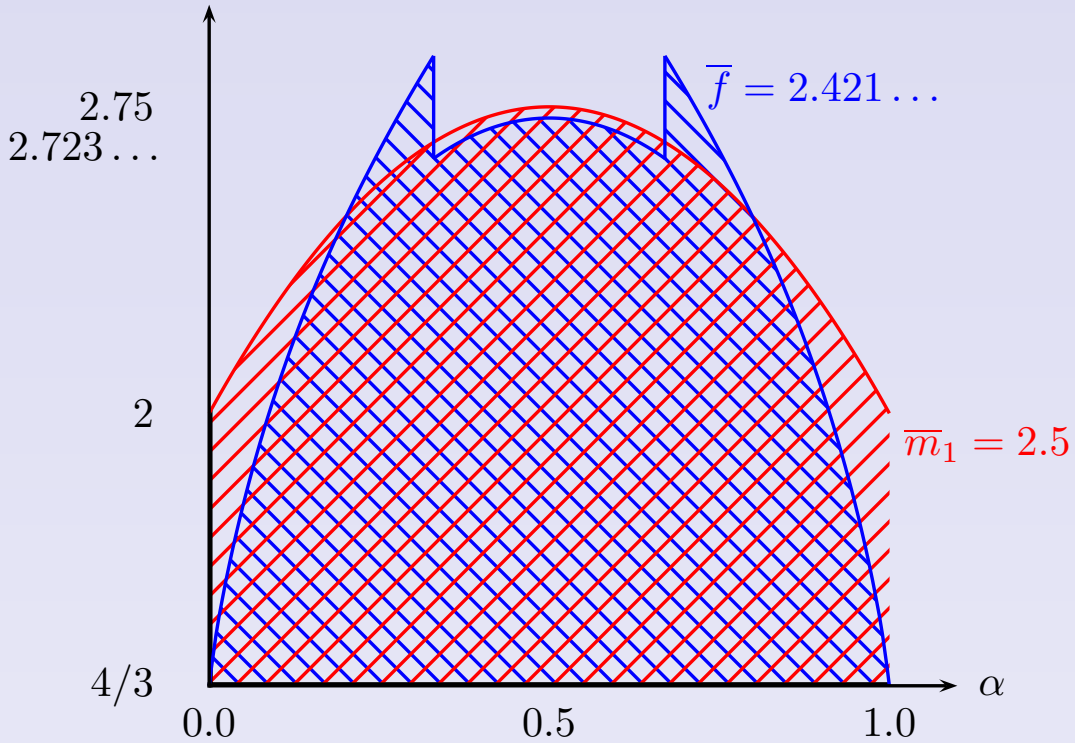
Proporcional-de-3: Batfind



Proporcional-de-3: Batfind



Proporcional-de-3: Batfind



ν-find

- Si $\alpha = 1/3 - \epsilon$ ó $\alpha = 2/3 + \epsilon$ entonces haríamos mejor usando la mediana de la muestra y no el menor o el mayor
- ν-find generaliza a batfind, situando los puntos de corte en $a_1 = \nu$ y $a_2 = 1 - \nu$
- La función $f_\nu(\alpha)$ característica de ν-find satisface las mismas ecuaciones diferenciales que la de batfind pero las constantes C_i dependen de ν: $C_i = C_i(\nu)$

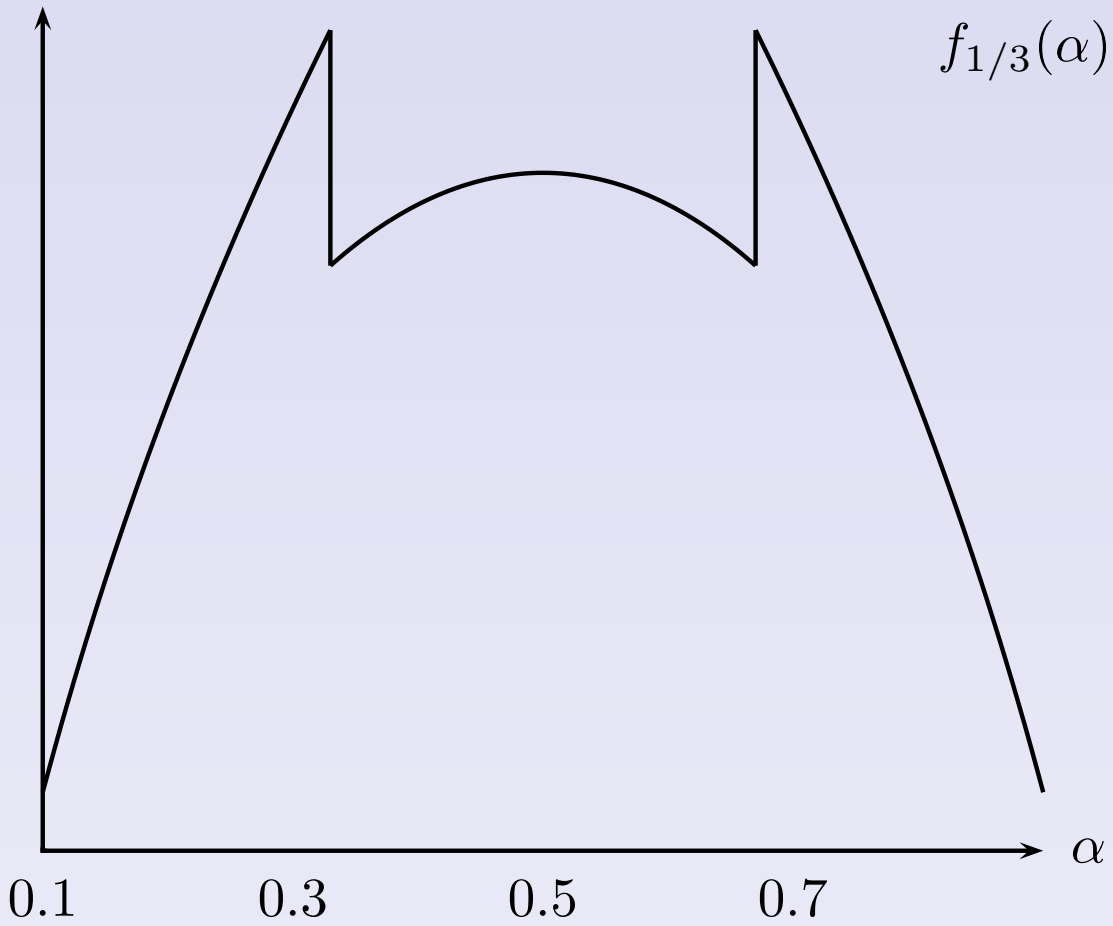
ν-find

- Si $\alpha = 1/3 - \epsilon$ ó $\alpha = 2/3 + \epsilon$ entonces haríamos mejor usando la mediana de la muestra y no el menor o el mayor
- ν-find generaliza a batfind, situando los puntos de corte en $a_1 = \nu$ y $a_2 = 1 - \nu$
- La función $f_\nu(\alpha)$ característica de ν-find satisface las mismas ecuaciones diferenciales que la de batfind pero las constantes C_i dependen de ν: $C_i = C_i(\nu)$

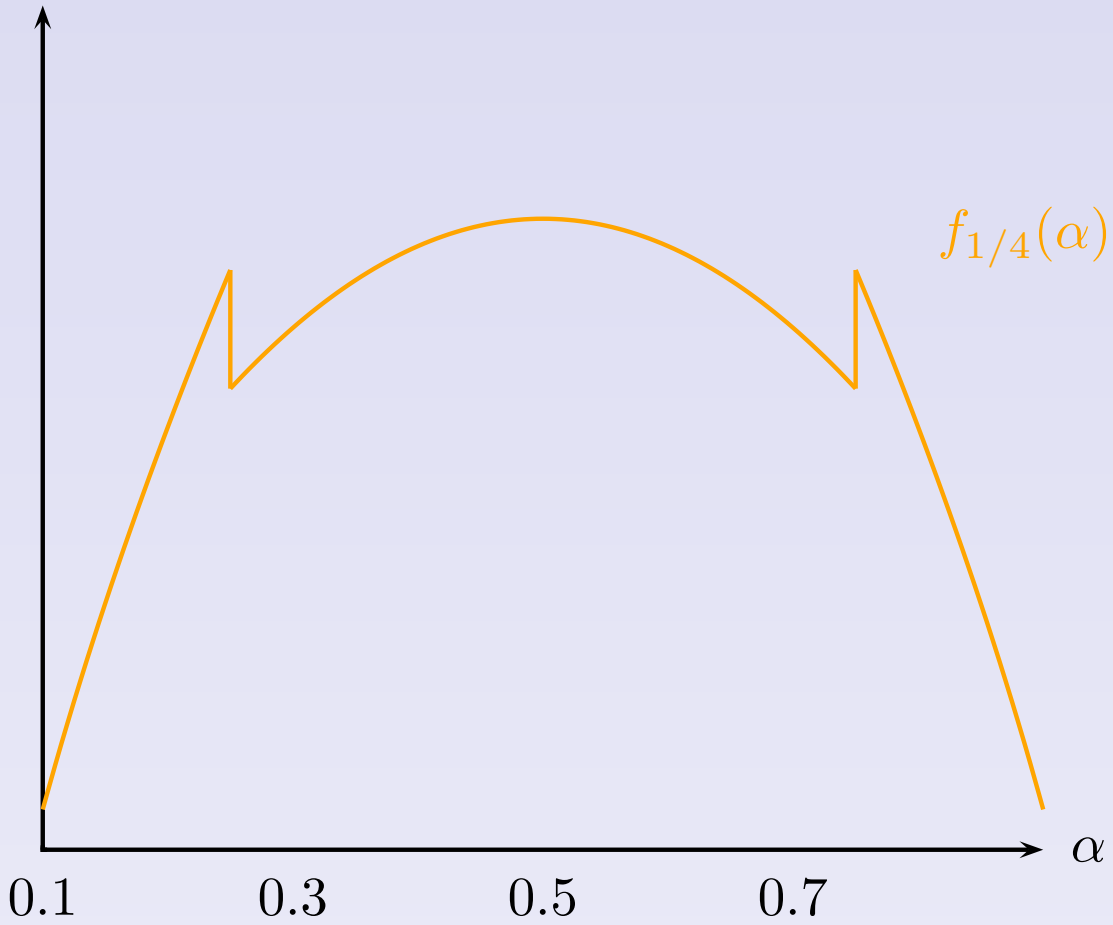
ν-find

- Si $\alpha = 1/3 - \epsilon$ ó $\alpha = 2/3 + \epsilon$ entonces haríamos mejor usando la mediana de la muestra y no el menor o el mayor
- ν-find generaliza a batfind, situando los puntos de corte en $a_1 = \nu$ y $a_2 = 1 - \nu$
- La función $f_\nu(\alpha)$ característica de ν-find satisface las mismas ecuaciones diferenciales que la de batfind pero las constantes C_i dependen de ν: $C_i = C_i(\nu)$

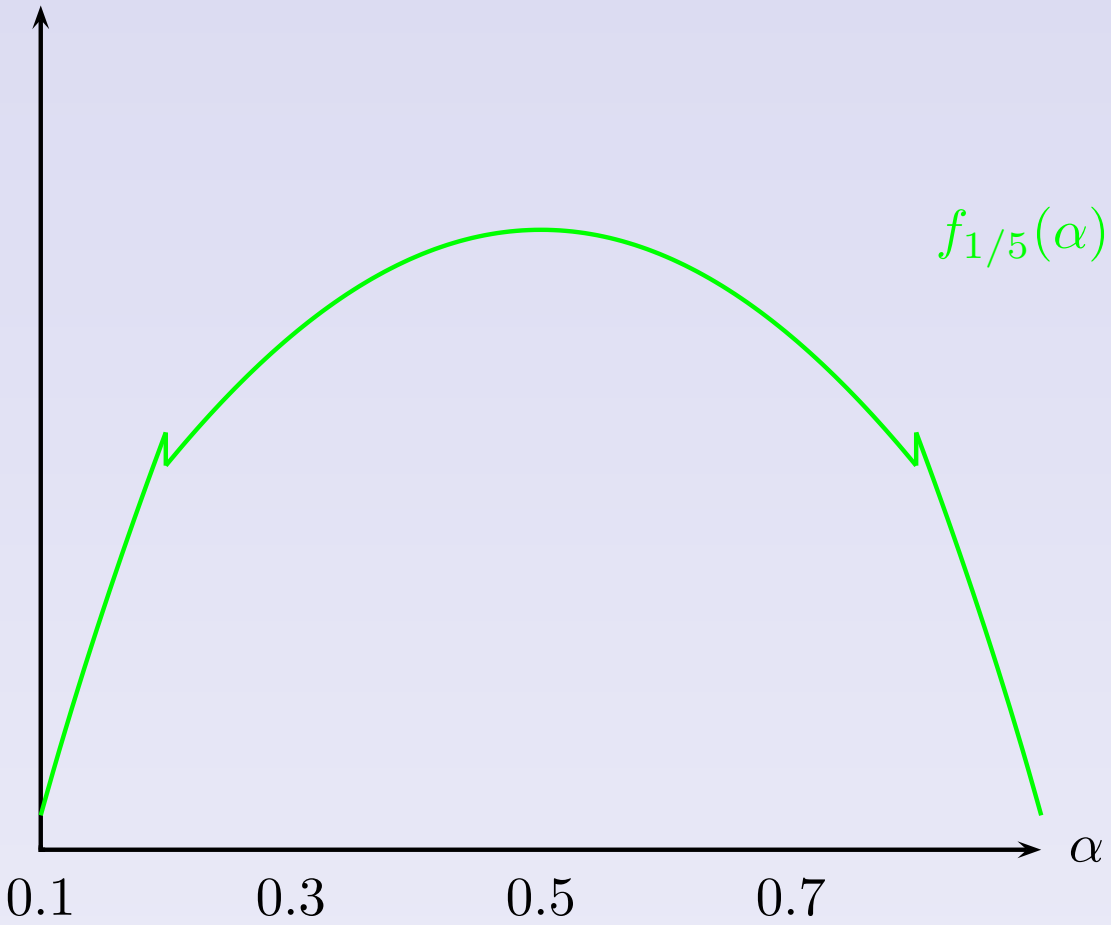
ν-find



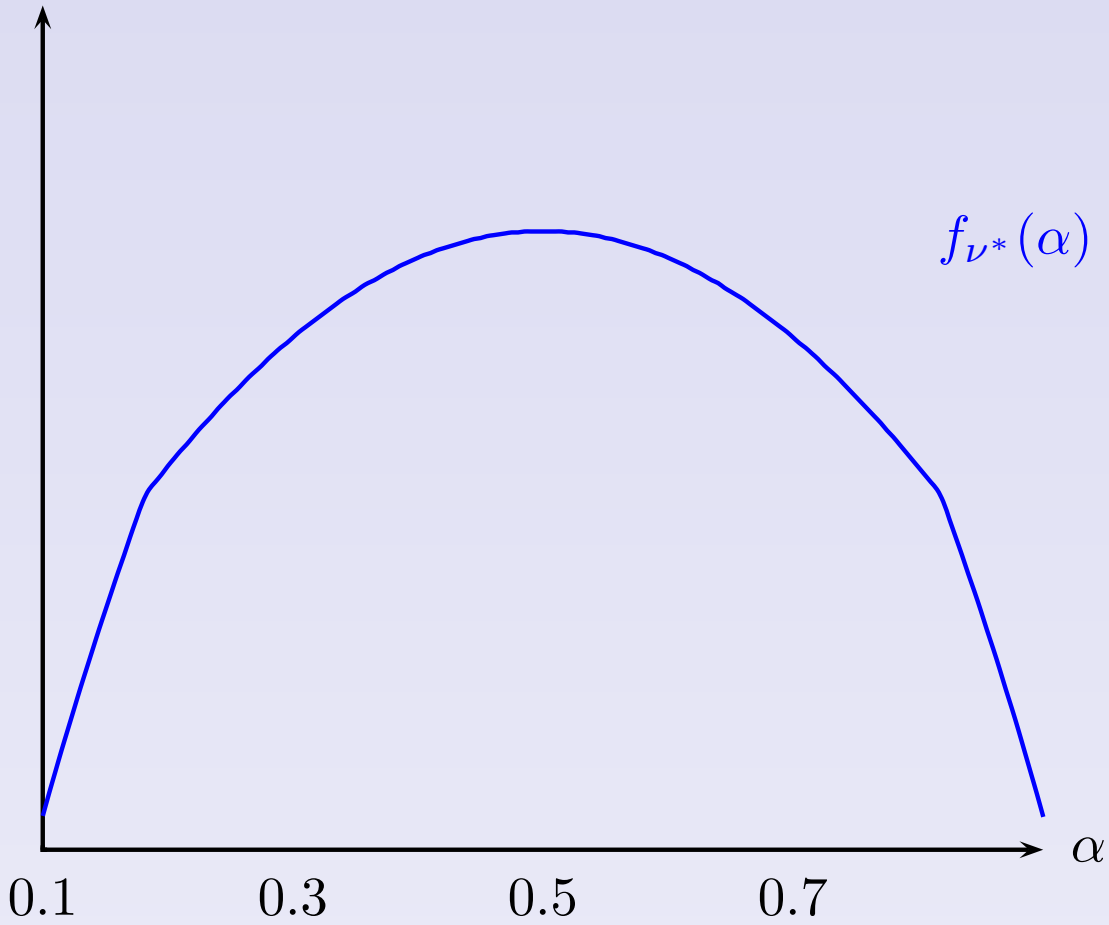
ν-find



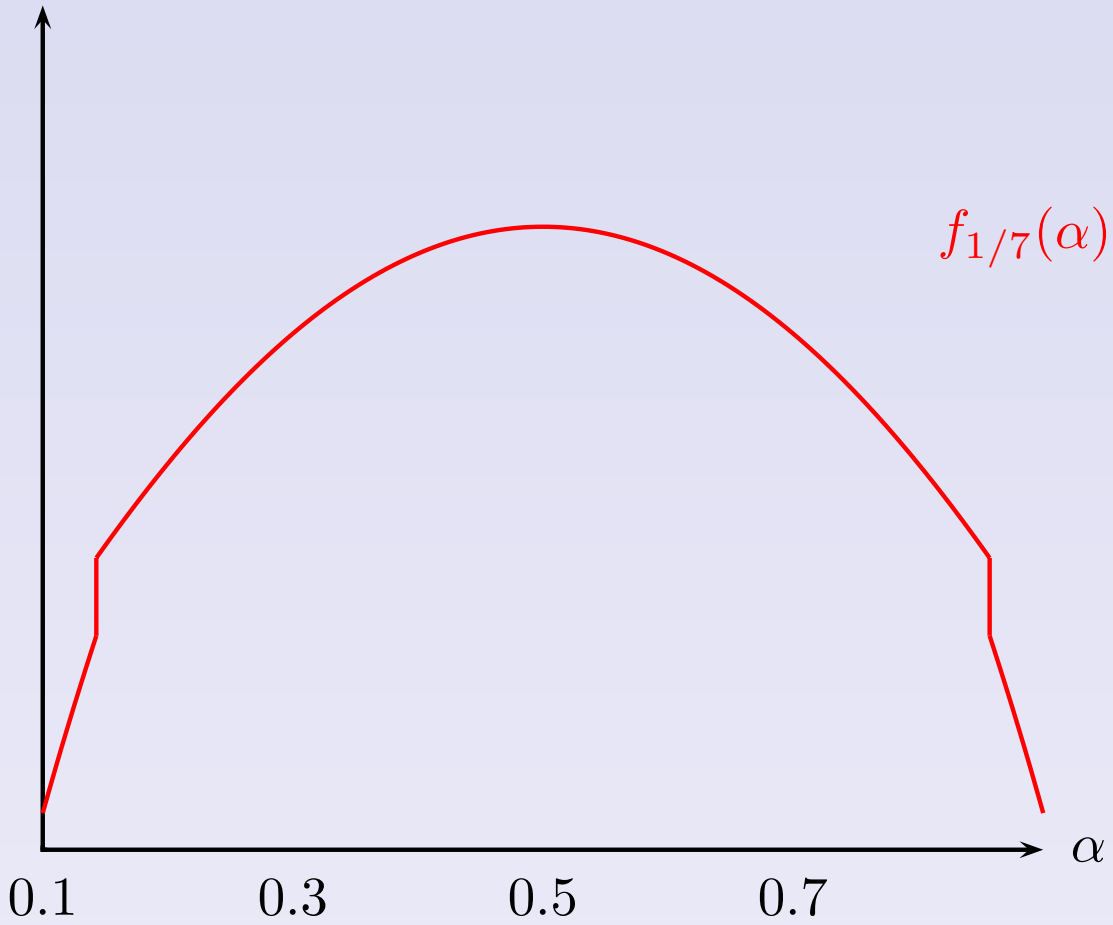
ν-find



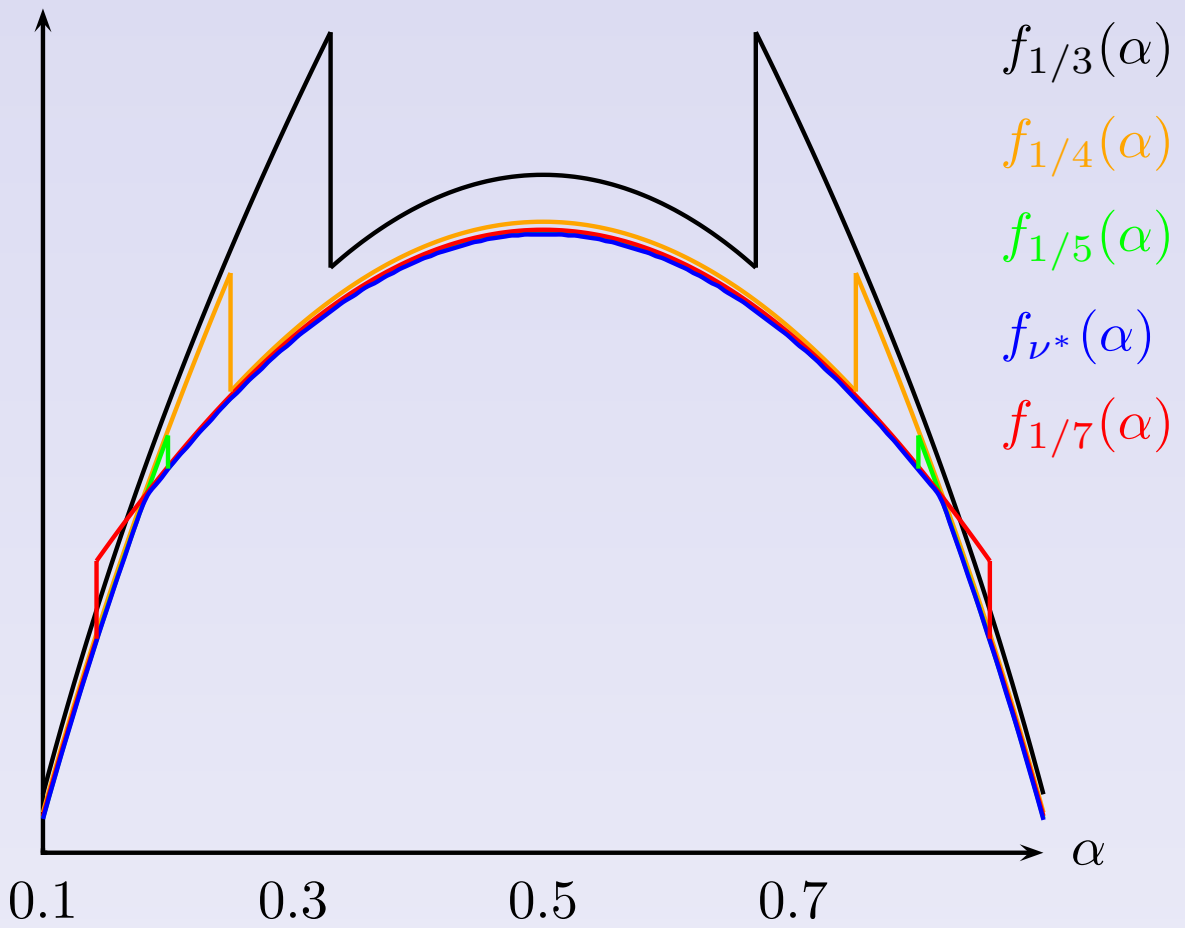
ν-find



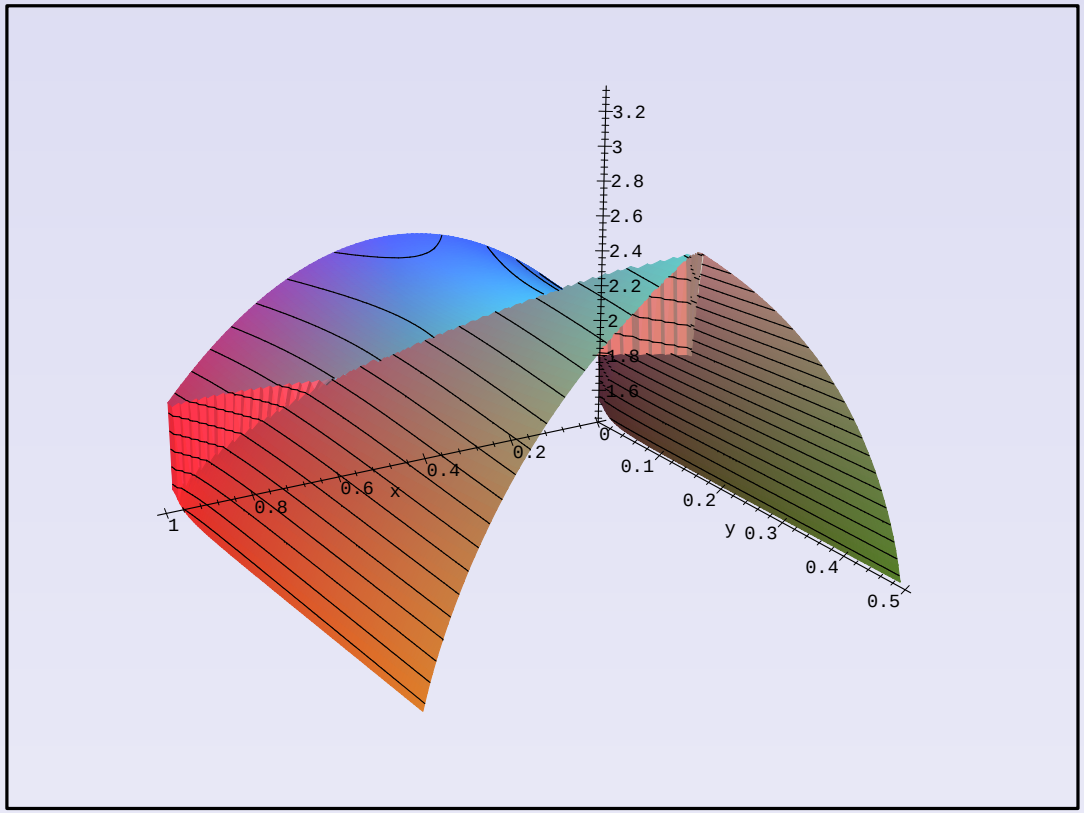
ν-find



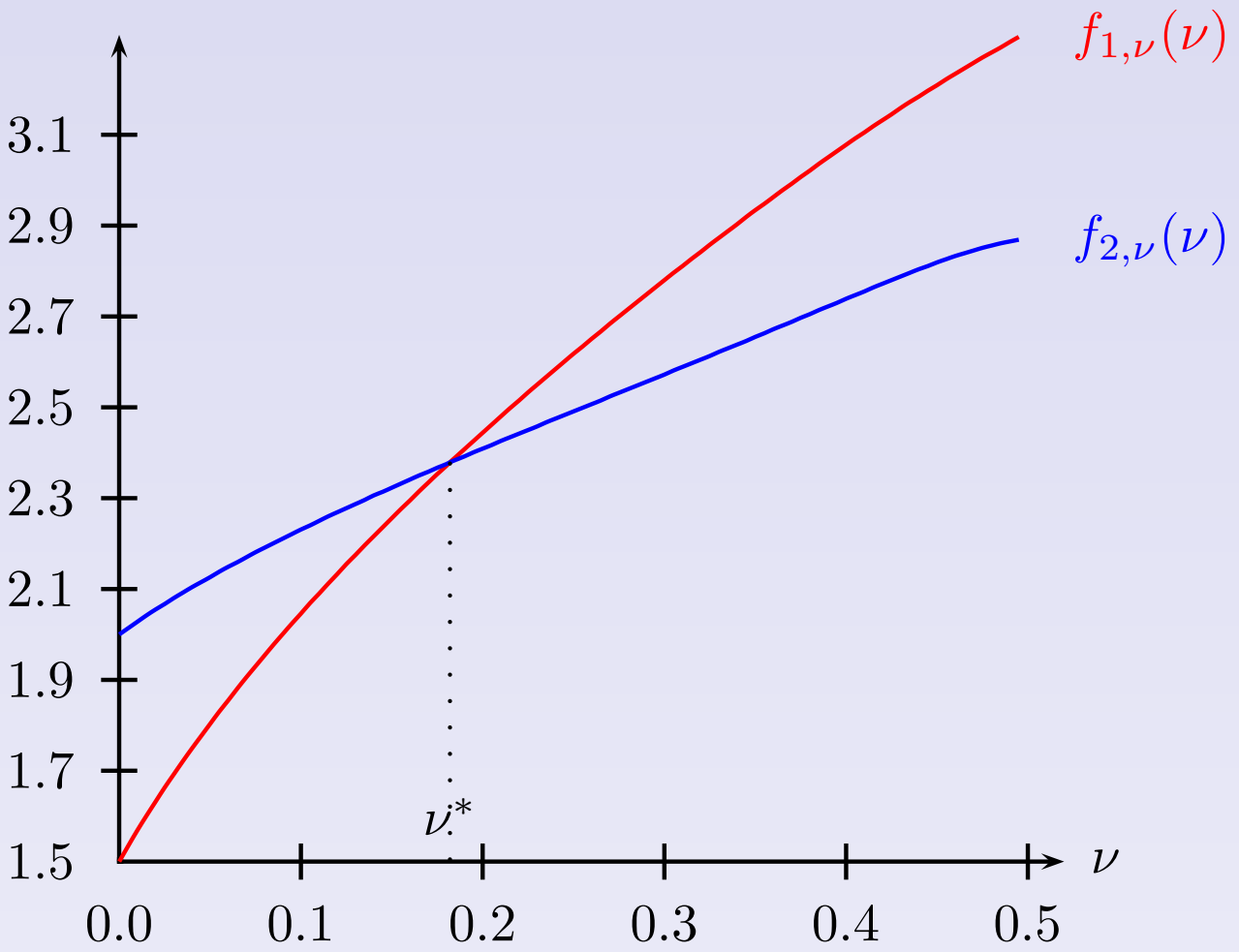
ν-find


$f_{1/3}(\alpha)$
$f_{1/4}(\alpha)$
$f_{1/5}(\alpha)$
$f_{\nu^*}(\alpha)$
$f_{1/7}(\alpha)$
α
0.1
0.3
0.5
0.7

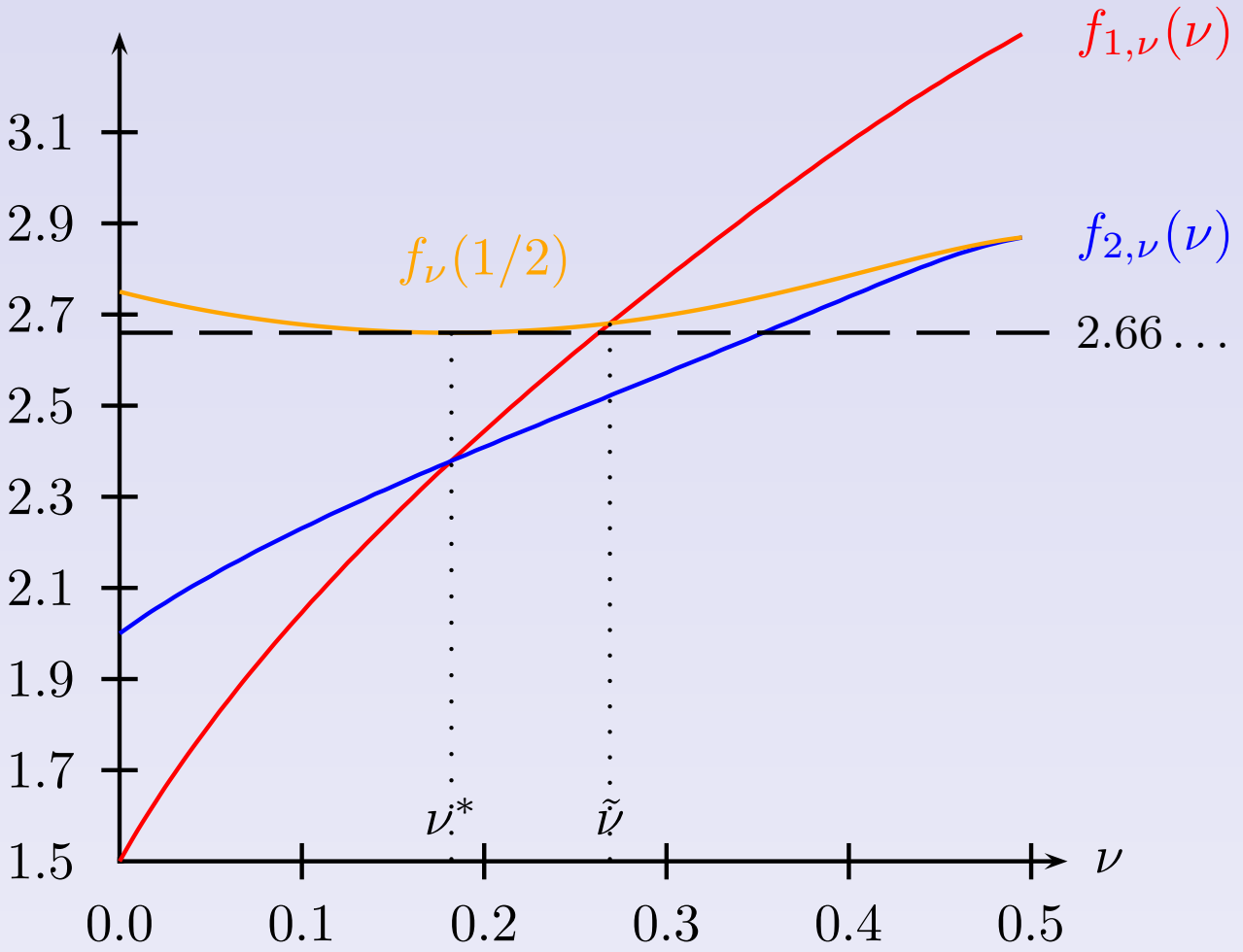
ν-find



ν-find



ν-find


$f_{1,\nu}(\nu)$
$f_{2,\nu}(\nu)$
$f_\nu(1/2)$
2.66 . . .
3.1
2.9
2.7
2.5
2.3
2.1
1.9
1.7
1.5
ν^*
$\tilde{\nu}$
ν
0.0
0.1
0.2
0.3
0.4
0.5

El valor óptimo de ν

Teorema

Existe un valor $\nu^* = 0{,}182\ldots$ *tal que para cualquier* ν, $0 < \nu < 1/2$, *y para cualquier* α, $0 \leq \alpha \leq 1$,

$$f_{\nu^*}(\alpha) \leq f_{\nu}(\alpha)$$

Además ν^* *es el único valor de* ν *en* $(0, 1/2)$ *tal que* f_ν *es continua, es decir,*

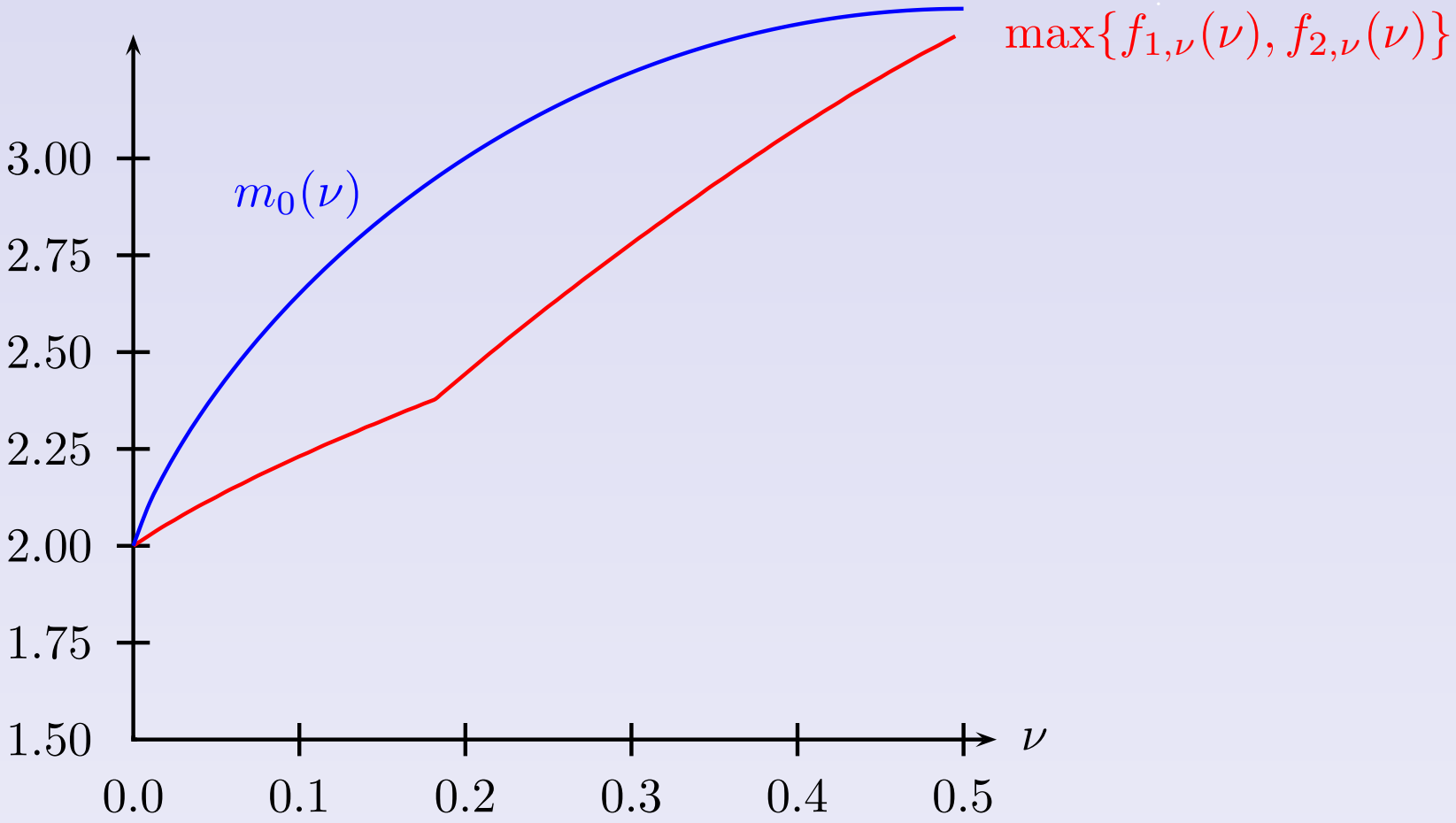
$$f_{\nu^*,1}(\nu^*) = f_{\nu^*,2}(\nu^*)$$

El valor óptimo de ν

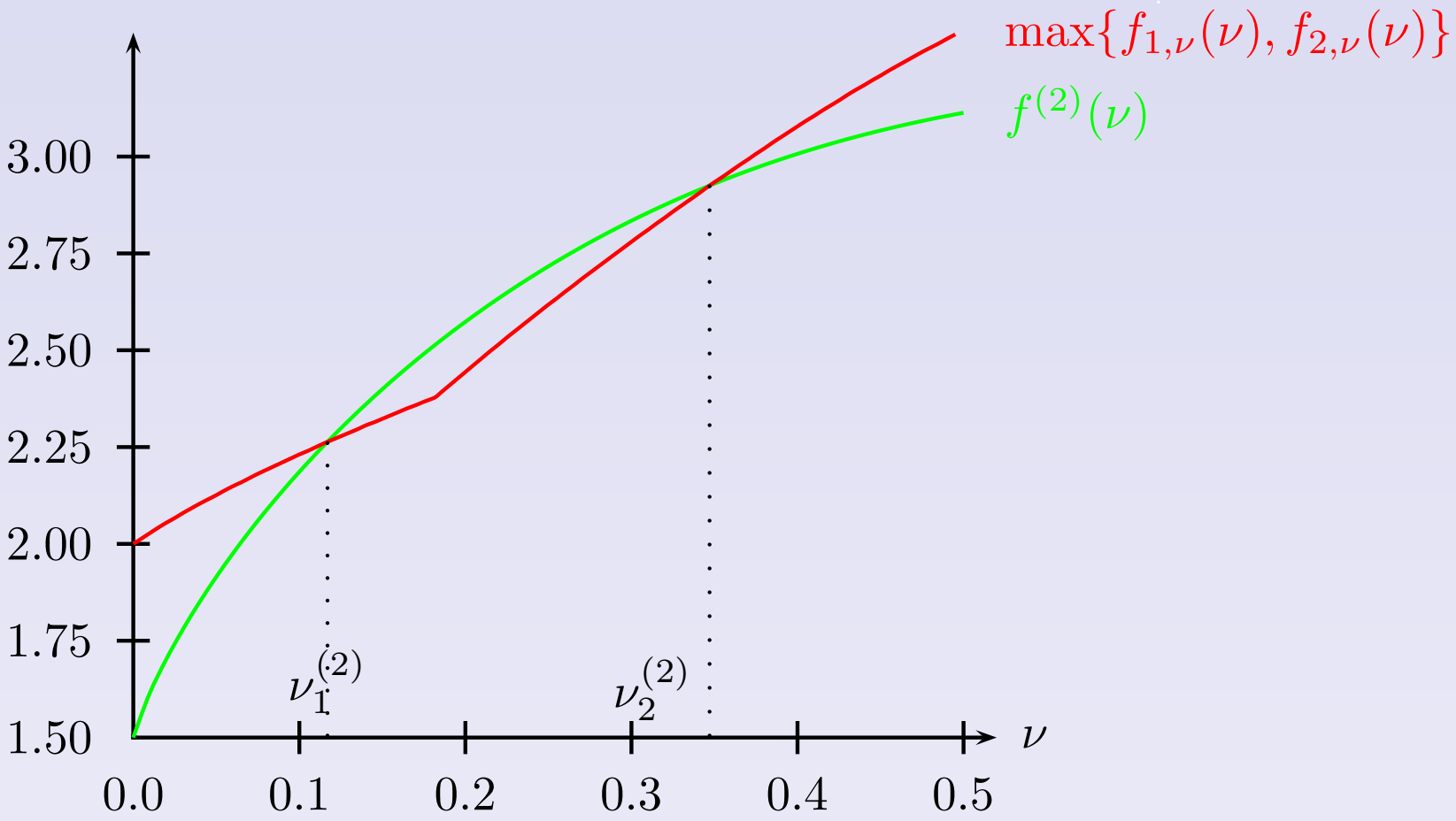
- El valor ν^* minimiza $f_\nu(1/2)$ (coste de localizar la mediana) y $\overline{f}_\nu$ (coste de localizar un rango aleatorio):

$$f_{\nu^*}(1/2) = 2{,}659\ldots$$
$$\overline{f}_{\nu^*} = 2{,}342\ldots$$

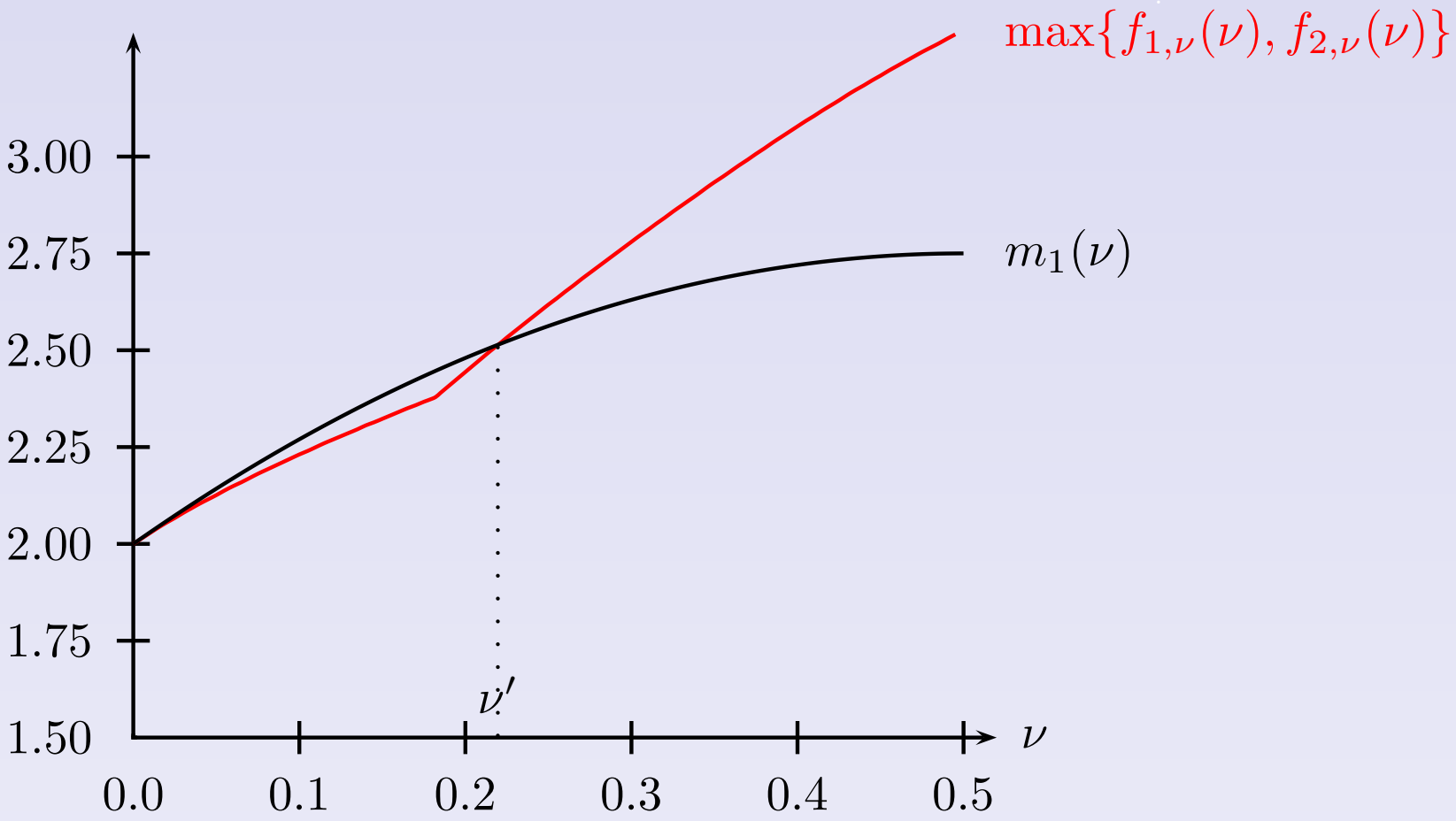
ν-find vs. quickselect estándar



ν-find vs. proporcional-de-2



ν-find vs. mediana-de-3



Optimalidad de las estrategias proporcionales

Teorema

Sea $\{f_s(\alpha)\}_{s\geq 1}$ *la secuencia de funciones características de una familia de estrategias simétricas con muestras de tamaño* s*. Si* $\lim_{s\to\infty} r(\alpha)/s = \alpha$ *(son de tipo proporcional) y* $r(\alpha) > \alpha \cdot s + 1 - \alpha$ *para* $\alpha \leq 1/2$ *y* s *suficientemente grande (no son proporcionales puras) entonces*

$$f_\infty(\alpha) = \lim_{s\to\infty} f_s(\alpha) = 1 + \min(\alpha, 1-\alpha)$$