Argmin and Monotonic Functions

① We've by now seen a few situations in which we want to compute $\underset{x}{\operatorname{argmin}}\ f(x)$ or $\underset{x}{\operatorname{argmax}}\ f(x)$:

e.g. $\underset{\theta}{\operatorname{argmin}}\ L(\theta)$ where L is a loss function

$\underset{w}{\operatorname{argmax}}\ P(w)\prod_{n=1}^{N} P(y^{(n)}|w, x^{(n)})$ for regression problems

In all of these situations, $f(x)$ has been nonnegative (i.e. $f: \mathbb{R} \to \mathbb{R}^+ \cup \{0\}$). Let's take a closer look at some strategies for computing $\underset{x}{\operatorname{argmin}}\ f(x)$ $\left[\text{or } \underset{x}{\operatorname{argmax}}\ f(x)\right]$, given that f has a nonnegative range.

---

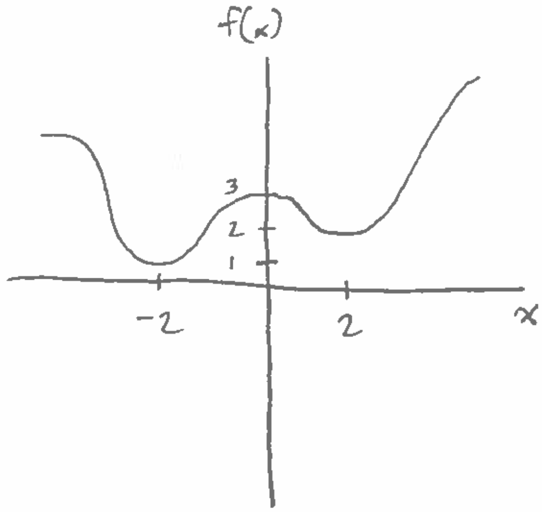
② One property we've already used (see Regression Problems ⑲) is:

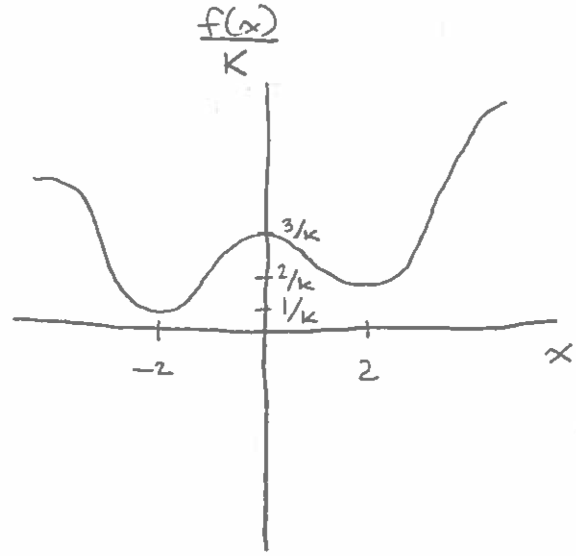
$$\underset{x}{\operatorname{argmin}}\ \frac{f(x)}{K} = \underset{x}{\operatorname{argmin}}\ f(x)$$

if K is a positive constant.

Argmin and Monotonic Functions

(3) This is easy enough to justify with a picture





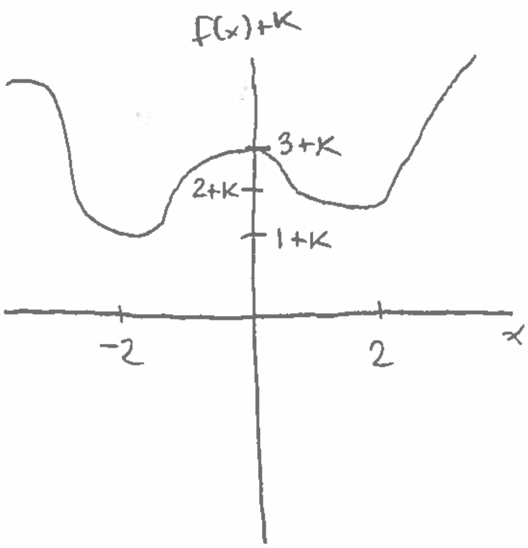
$\underset{x}{\operatorname{argmin}}\ f(x) = -2$

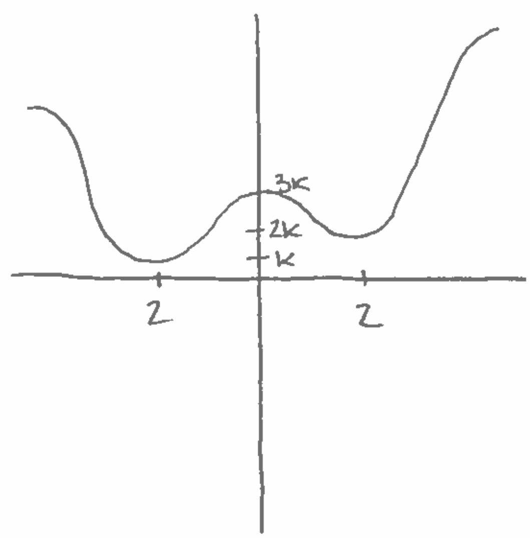
$\underset{x}{\operatorname{argmin}}\ \frac{f(x)}{K} = -2$

(4) The same trick works in several other varieties:

$$\underset{x}{\operatorname{argmin}}\ f(x) + K = \underset{x}{\operatorname{argmin}}\ f(x) \quad \text{for any } K \in \mathbb{R}$$

$$\underset{x}{\operatorname{argmin}}\ K f(x) = \underset{x}{\operatorname{argmin}}\ f(x) \quad \text{for positive } K$$





$\underset{x}{\operatorname{argmin}}\ f(x) + K = -2$

$\underset{x}{\operatorname{argmin}}\ K \cdot f(x) = -2$

ARGMIN AND MONOTONIC FUNCTIONS

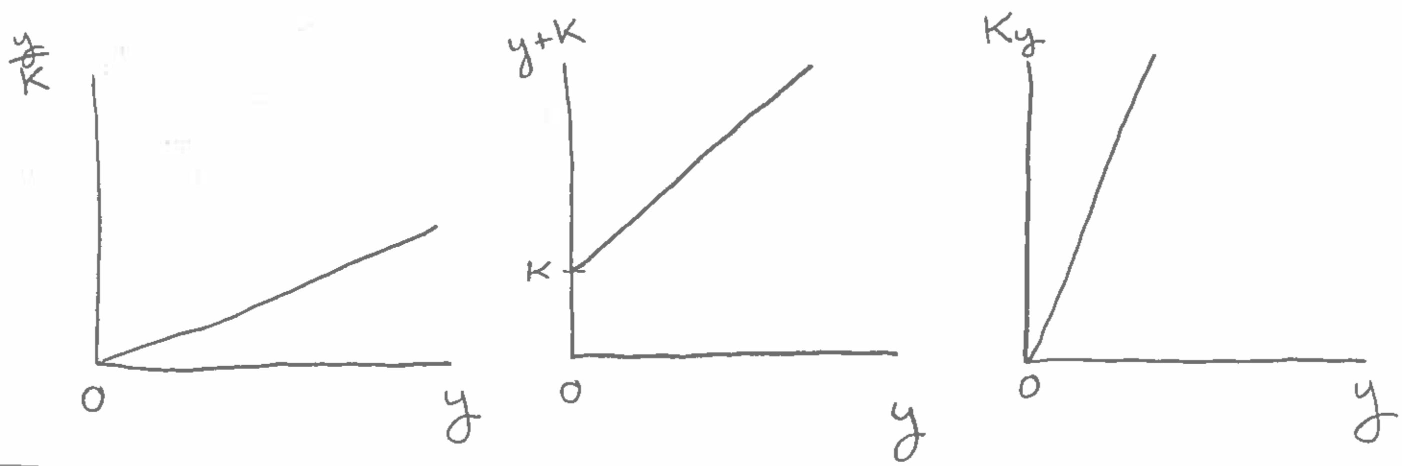
5) How can we generalize this trick? Well, let's look at these transformations as functions of $f(x)$, i.e.

$$g(y) = \frac{y}{K} \Rightarrow g(f(x)) = \frac{f(x)}{K}$$

$$g(y) \Rightarrow y + K \Rightarrow g(f(x)) = f(x) + K$$

$$g(y) \Rightarrow Ky \Rightarrow g(f(x)) = Kf(x)$$

We've assumed $f(x) \geq 0$ for all x, so let's look at $g(y)$ over the domain $[0, \infty)$:



6) All are monotonically increasing, i.e.

$$y_1 > y_2 \Leftrightarrow g(y_1) > g(y_2) \quad \text{for } y_1, y_2 \in \mathbb{R}^+ \cup \{0\}$$

This means:

$$f(x_1) > f(x_2) \Leftrightarrow g(f(x_1)) > g(f(x_2)) \quad \text{for } x_1, x_2 \in \mathbb{R}$$

which implies

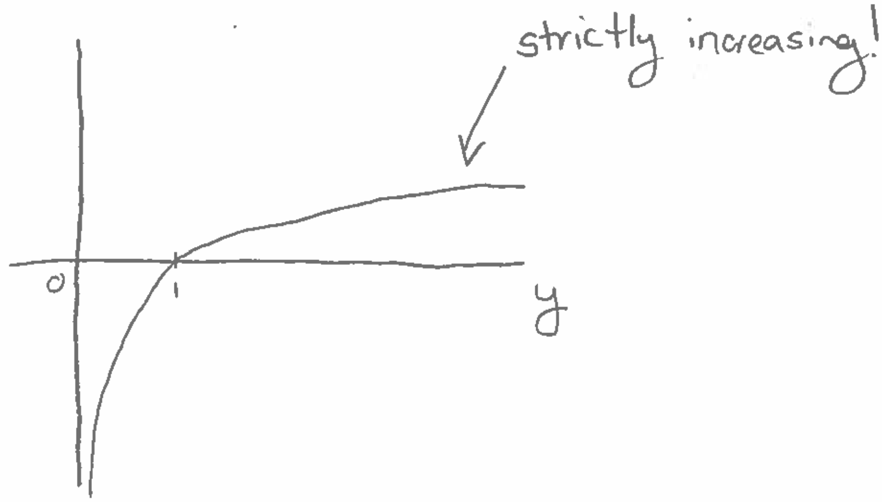
$$\underset{x}{\operatorname{argmin}}\ f(x) = \underset{x}{\operatorname{argmin}}\ g(f(x))$$

[Exercise: prove by contradiction]

ARGMIN AND MONOTONIC FUNCTIONS

⑦ The uncontested monarch of monotonic functions (over $\mathbb{R}^+$) is the logarithm, i.e.

$$g(y) = \log y$$



Because it is monotonically increasing, we know from ⑥ that:

$$\underset{x}{\operatorname{argmin}}\ f(x) = \underset{x}{\operatorname{argmin}}\ \log f(x)$$

⑧ This can be amazingly convenient when $f(x)$ is the product of complicated functions, e.g. suppose $f(x) = (2 + \sin x)(2 + \cos x)(x^2 + 1)$. First we observe that the range of $f(x)$ is positive. Thus:

$$\begin{aligned}
\underset{x}{\operatorname{argmin}}\ f(x) &= \underset{x}{\operatorname{argmin}}\ \log f(x) \\
&= \underset{x}{\operatorname{argmin}}\ \log (2 + \sin x)(2 + \cos x)(x^2 + 1) \\
&= \underset{x}{\operatorname{argmin}}\ \log (2 + \sin x) + \log (2 + \cos x) + \log (x^2 + 1)
\end{aligned}$$

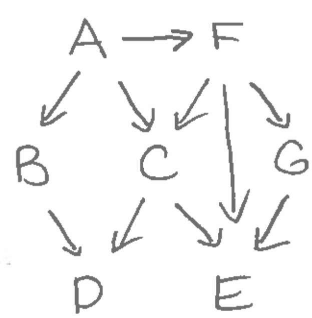
ARGMIN AND MONOTONIC FUNCTIONS

⑨ The product has become a sum! Thus we can easily take the derivative:

$$\frac{d}{dx} \log f(x) = \frac{\cos x}{2+\sin x} - \frac{\sin x}{2+\cos x} + \frac{2x}{x^2+1}$$

---

⑩ This trick can be particularly useful when dealing with joint probability distributions. Consider a distribution that factors according to the following Bayesian network:



i.e. $P(a,b,c,d,e,f,g) = P(a)P(b|a)P(c|a,f)P(d|b,c)$
$\cdot P(e|c,f,g)P(f|a)P(g|f)$

Suppose we wanted to know the value of F that was most likely, given the other variables:

$$\hat{f} = \arg\max_f P(f|a,b,c,d,e,g)$$

$$= \arg\max_f \frac{P(a,b,c,d,e,f,g)}{\sum_f P(a,b,c,d,e,f,g)}$$

ARGMIN AND MONOTONIC FUNCTIONS

(11) First thing to notice is that $P(a,b,c,d,e,g)$ is the same no matter what value we choose for f, so we can say:

$$\hat{f} = \operatorname*{argmax}_f \frac{P(a,b,c,d,e,f,g)}{K}$$

$$= \operatorname*{argmax}_f P(a,b,c,d,e,f,g) \quad [\text{assuming } K>0]$$

(12) We can then use the Bayesian network to simplify further:

$$\hat{f} = \operatorname*{argmax}_f P(a)P(b|a)P(c|a,f)P(d|b,c)P(e|c,f,g)P(f|a)P(g|f)$$

There are two issues with this formula:

- products are tough to differentiate
- the overall quantity gets very small, very fast. Suppose each conditional probability term is .01 (i.e. a 1% probability). Then the overall expression evaluates to $(.01)^7 = .00000000000001$. Pretty quickly this gets too small to represent in memory (underflow).

(13) But since $\log y$ is monotonically increasing over $\mathbb{R}^+$:

$$\hat{f} = \underset{f}{\operatorname{argmax}} \log P(a)P(b|a)P(c|a,f)P(d|b,c)P(e|c,f,g)P(f|a)P(g|f)$$

$$= \underset{f}{\operatorname{argmax}} \log P(a) + \log P(b|a) + \log P(c|a,f) + \log P(d|b,c) + \log P(e|c,f,g) + \log P(f|a) + \log P(g|f)$$

Now, not only do we have an easy sum to differentiate, the quantity is also easy to represent in memory:

$$7 \cdot \log(.01) = 7 \cdot (-2)$$
$$= -14$$