OPTIMIZATION FOR MACHINE LEARNING

September 19, 2024

Today: Intro to optimization (2/2), with exercises

Coming up: Sep. 23: Gradient descent (Gabriel Peyré)
Sep 26 : LAB on gradient descent (C. Royer)

Back where we left off

"Find a C^1 function f such that f is not convex

$$\nabla f(\bar{x}) = 0 \iff \bar{x} \in \operatorname{argmin}_x f(x)$$ "

NB: Such functions are called invex functions (or, more rarely, pseudo-convex functions)

$$\{\text{invex functions}\} \supset \{\text{convex functions}\}$$

In 1 dimension

$$x \longmapsto x^2 + 2\sin(x)^2$$

$$\varphi: x \longmapsto -e^{-x^2} \qquad \varphi'(x) = 2x e^{-x^2}$$

$= 0$ iff $x = 0$

In 2 dimensions:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \longmapsto x_1^2 - e^{-x_2^2}$$

One application of invex functions

Data: $\{(a_i, y_i)\}_{i=1..m}$ $\quad a_i \in \mathbb{R}^d \quad y_i \in \{0,1\}$

One optimization problem for binary classification:

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{2m} \|\sigma(Ax) - y\|^2$$

$$= \frac{1}{2m} \sum_{i=1}^{m} \left(\sigma(a_i^T x) - y_i\right)^2$$

ℓ_2 loss, model $a \longmapsto \sigma(a^T x) = \frac{1}{1+e^{-a^T x}}$

(sigmoid function)

Regularization:

1) ℓ_2 regularization

$$\min_{x \in \mathbb{R}^d} f(x) + \underbrace{\frac{\lambda}{2}\|x\|^2}_{\text{convex}}$$

· For sufficiently large λ, we expect $f + \frac{\lambda}{2}\|\cdot\|^2$ to behave (almost) like a convex function

2) Invex regularization

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{2n}\|\sigma(Ax) - y\|^2$$

$$\min_{\substack{x \in \mathbb{R}^d \\ p \in \mathbb{R}^n}} \hat{f}(x,p) = \frac{1}{2n}\|\sigma(Ax) - y + \lambda p\|^2$$

$$= \frac{1}{2n}\|\sigma(Ax) - y\|^2 + \frac{\lambda^2}{2n}\|p\|^2 + \frac{\lambda}{n} p^T(\sigma(Ax) - y)$$

The function $\hat{f} : \mathbb{R}^{d+n} \longrightarrow \mathbb{R}$ is invex $\forall \lambda > 0$

Final note about invex functions

If $f : \mathbb{R}^d \to \mathbb{R}$ is C^1 and invex, then there exists a C^1 function $\eta : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}^d$ such that

$$\forall (x,y) \in (\mathbb{R}^d)^2, \quad f(y) \geq f(x) + \nabla f(x)^T \eta(x,y)$$

$\Rightarrow \eta(x,y) = y - x$, recover C^1 convex functions

$\left(f\ C^1 \text{ convex} \iff f(y) \geq f(x) + \nabla f(x)^T(y-x) \quad \forall (x,y) \in \mathbb{R}^d \right)$

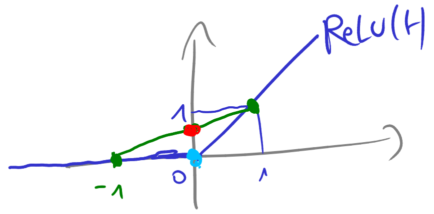
MORE ON CONVEX FUNCTIONS

↳ We defined convexity only for C^1 functions using the gradient
=> the notion of a convex function does not involve derivatives

Def: A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if

$$\forall (x,y) \in (\mathbb{R}^d)^2,\ \forall \alpha \in [0,1],\ f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$$

Ex) ReLU : $t \mapsto \max(t, 0)$



$x = -1$
$y = 1 \quad \alpha = \frac{1}{2}$

$f(\alpha x + (1-\alpha) y) = f(0) = 0$
$\alpha f(x) + (1-\alpha) f(y) = \frac{1}{2} \times 0 + \frac{1}{2} \times 1 = \frac{1}{2}$

↳ Checking convexity for functions with more than 2 variables is difficult !
=> We use a set of rules to certify convexity of most convex functions used in ML

Rules:

I. $f : \mathbb{R}^d \to \mathbb{R}^m$, $x \mapsto Wx + b$, $W \in \mathbb{R}^{m \times d}$, $b \in \mathbb{R}^m$ is convex

II. All norms in $\mathbb{R}^d$ are convex

$\left(\sqrt{\sum_{i=1}^{d} x_i^2}\right) = \|\cdot\|$, $\|x\|_1 \left(= \sum_{i=1}^{d} |x_i|\right)$, $\|x\|_\infty \left(= \max_{1 \le i \le d} |x_i|\right)$

III. For any convex function f and any $\alpha \geq 0$, αf is convex

IV. f, g convex $\Rightarrow f+g$ convex

V. f, g convex $\Rightarrow \max(f,g)$ convex

VI. $f : \mathbb{R}^m \to \mathbb{R}$ convex $\Rightarrow x \mapsto f(Wx+b)$ convex
$W \in \mathbb{R}^{m \times d}$, $b \in \mathbb{R}^m$

Proof of V:

$\forall (x,y) \in (\mathbb{R}^d)^2$, $\forall \alpha \in [0,1]$,

$$\max\left(f(\alpha x + (1-\alpha)y), g(\alpha x + (1-\alpha)y)\right)$$

If the max is attained at f

$$\max\left(f(\alpha x + (1-\alpha)y), g(\alpha x + (1-\alpha)y)\right)$$

$$= f(\alpha x + (1-\alpha)y) \overset{\text{convexity of } f}{\le} \alpha f(x) + (1-\alpha) f(y)$$

$$\le \underbrace{\alpha \max(f(x), g(x)) + (1-\alpha)\max(f(y), g(y))}_{\text{convexity inequality}}$$

The case where the maximum is attained for g is handled the same way.

Exercise: Show that the objective function for SVM

$$x \mapsto \frac{1}{m} \sum_{i=1}^{m} \max(1 - y_i a_i^T x, 0)$$ is convex (using the rules)

$(a_i \in \mathbb{R}^d, y_i \in \{-1,1\}, x \in \mathbb{R}^d)$

Solution: By I, $x \mapsto 0$ is convex. and $x \mapsto 1 - y_i a_i^T x$ is convex $\forall i$

By V, $\forall i$, $x \mapsto \max(1 - y_i a_i^T x, 0)$

By IV (applied $n-1$ times), $x \mapsto \sum_{i=1}^{n} \max(1 - y_i a_i^T x, 0)$ is convex

By III, $x \mapsto \frac{1}{n} \sum_{i=1}^{n} \max(1 - y_i a_i^T x, 0)$ is convex

↳ Another characterization of convexity for C^2 functions

Def.: $f : \mathbb{R}^d \to \mathbb{R}$ is C^2 (twice continuously differentiable) if it is C^1 and $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ is also C^1

f C^2 $\Rightarrow$ $\exists\, \nabla^2 f(x) \in \mathbb{R}^{d \times d}$ symmetric matrix such that

2nd-order Taylor expansion of f around x

$$f(y) \simeq f(x) + \nabla f(x)^T (y-x) + \frac{1}{2}(y-x)\nabla^2 f(x)(y-x)$$

when $\|y-x\|$ is small

Note: $\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_d}(x) \end{bmatrix}$ and $\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_d}(x) & \cdots & \frac{\partial^2 f}{\partial x_d^2}(x) \end{bmatrix}$

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x) = \frac{\partial^2 f}{\partial x_j \partial x_i}(x)$$

$\nabla^2 f(x)$: Hessian matrix

Theorem: Let $f : \mathbb{R}^d \to \mathbb{R}$ be C^2.

$$\left[f \text{ is convex} \right] \iff \left[\forall x \in \mathbb{R}^d,\ \nabla^2 f(x) \succeq 0 \right]$$

"positive semidefinite"

$\succeq$: $A \in \mathbb{R}^{d \times d}$ is positive semidefinite (or PSD)
if $A = A^T$ (A symmetric) and

$$\forall x \in \mathbb{R}^d,\ \underbrace{\underbrace{x^T}_{1\times d} \underbrace{\overset{d\times d}{A} \overset{d\times 1}{x}}_{d\times 1}}_{1\times 1} \geq 0$$

$M \in \mathbb{R}^{d \times m}$ $\quad M = [M_{ij}]_{\substack{1 \le i \le d \\ 1 \le j \le m}}$

$M^T \in \mathbb{R}^{m \times d}$ $\quad [M^T]_{ji} = M_{ij}$ $\qquad x = \square \qquad x^T = \square$

Ex)

$A \in \mathbb{R}^{m \times d}$

$y \in \mathbb{R}^m$

$x \in \mathbb{R}^d$

- $f(x) = \frac{1}{2m} \|Ax - y\|^2$ linear least squares regression

$$\nabla f(x) = \frac{1}{m} A^T (Ax - y) \in \mathbb{R}^d$$

$$\nabla^2 f(x) = \frac{1}{m} A^T A$$

$$\forall v \in \mathbb{R}^d,\ v^T \nabla^2 f(x) v = v^T \left(\frac{A^T A}{m} \right) v = \frac{1}{m} \overbrace{v^T A^T}^{(Av)^T} A v$$

$$= \frac{1}{m} \|Av\|^2 \geq 0$$

- $f(x) = \frac{1}{m} \sum_{i=1}^m \log\left(1 + \exp(-y_i a_i^T x)\right)$

logistic loss $\qquad a_i \in \mathbb{R}^d,\ y_i \in \{0, 1\}$

$$\Rightarrow \forall i = 1..m,\ f_i(x) = \log\left(1 + \exp(-y_i a_i^T x)\right)$$

$$\frac{-y_i \exp(-y_i a_i^T x)}{1+\exp(-y_i a_i^T x)} = \nabla f_i(x) = \overbrace{\frac{-y_i}{1+\exp(y_i a_i^T x)}}^{\in\mathbb{R}} \overbrace{a_i}^{\in\mathbb{R}^d}$$

$$\nabla^2 f_i(x) = \underbrace{\frac{y_i^2 \exp(y_i a_i^T x)}{(1+\exp(y_i a_i^T x))^2}}_{\in\mathbb{R}} \overbrace{\underbrace{a_i}_{d\times 1}\underbrace{a_i^T}_{1\times d}}^{d\times d}$$

NB: $(a_i a_i^T)^T = (a_i^T)^T a_i^T$
$= a_i a_i^T$

$$\forall v \in \mathbb{R}^d, \quad v^T \nabla^2 f_i(x) v = \frac{y_i^2 \exp(y_i a_i^T x)}{(1+\exp(y_i a_i^T x))^2} \underbrace{v^T}_{1\times d} \underbrace{a_i a_i^T}_{d\times d} \underbrace{v}_{d\times 1}$$

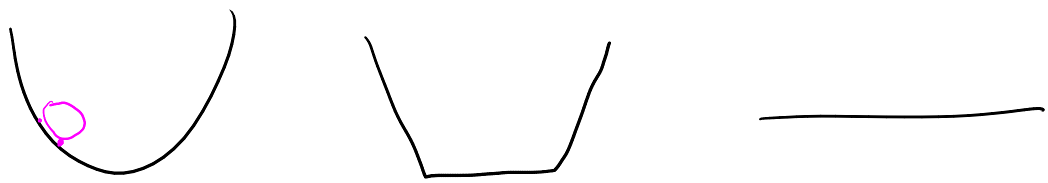
$$= \frac{y_i^2 \exp(y_i a_i^T x)}{(1+\exp(y_i a_i^T x))^2} \cdot \underbrace{a_i^T v}_{1\times 1} \underbrace{v^T a_i}_{1\times 1}$$

$$= \frac{y_i^2 \exp(y_i a_i^T x)}{(1+\exp(y_i a_i^T x))^2} (a_i^T v)^2 \geq 0$$

↳ <u>Convexity</u>:

→ can be checked for C^2, C^1, and even not C^1 functions
(even true for functions $\mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$)

→ for convex functions, $\|\nabla f(\bar{x})\| = 0$
$\iff \bar{x} \in \operatorname{argmin}_{x \in \mathbb{R}^d} f(x)$



Convex   Convex   Convex

Def: (Strongly convex function)

- $f: \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex $(\mu > 0)$ if
  $\forall (x,y) \in (\mathbb{R}^d)^2$, $f(\alpha x + (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) f(y) - \frac{\mu \alpha (1-\alpha)}{2} \|y-x\|^2$
- $f: \mathbb{R}^d \to \mathbb{R}$ C^1 is μ-strongly convex $(\mu > 0)$ if
  $\forall (x,y) \in (\mathbb{R}^d)^2$, $f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2} \|y-x\|^2$
- $f: \mathbb{R}^d \to \mathbb{R}$ C^2 is μ-strongly convex $(\mu > 0)$ if

$$\nabla^2 f(x) \succeq \mu I$$

$\left(\Leftrightarrow \nabla^2 f(x) - \mu I \succeq 0 \right)$

$I = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix} \in \mathbb{R}^{d \times d}$

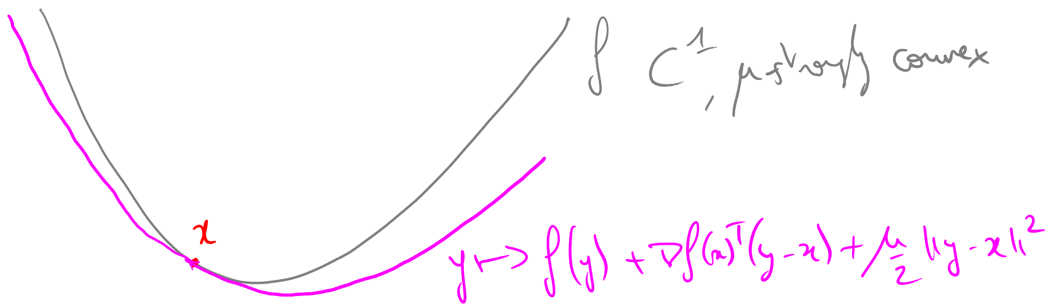
- Strongly convex functions are convex
- If f is μ-strongly convex, then $g: x \mapsto f(x) - \frac{\mu}{2}\|x\|^2$ is convex

$$\nabla^2 g(x) = \nabla^2 f(x) - \mu I \succeq 0$$

$\varphi: x \mapsto \frac{\mu}{2}\|x\|^2$ $\qquad$ $\nabla \varphi(x) = \mu x$

$\nabla^2 \varphi(x) = \mu I$

- Examples of strongly convex functions
  - $x \mapsto \frac{\mu}{2}\|x\|^2$
  - If f is convex, $f + \frac{\mu}{2}\|x\|^2$ is μ-strongly convex



Properties of strongly convex functions for optimization

- If f is strongly convex, it has at most one global minimum, i.e. $\left|\arg\min_x f(x)\right| \leq 1$.
- If f is strongly convex and continuous on $\mathbb{R}^d$, then it has a unique minimum,
- If f is strongly convex and C^1 on $\mathbb{R}^d$, then it has a unique minimum and that minimum is the only solution to $\nabla f(x) = 0$

NB: Strict convexity $\neq$ μ-strong convexity

f C^2 $\quad f$ strictly convex $\iff \nabla^2 f(x) \succ 0 \quad \forall x \in \mathbb{R}^d$

$$\iff v^T \nabla^2 f(x) v > 0 \quad \forall x \in \mathbb{R}^d, \ \forall v \in \mathbb{R}^d \setminus \{0\}$$

Ex) $x \longmapsto e^{-x}$

strictly convex but not strongly convex

$\nabla^2 f(x) = e^{-x} > 0 \quad \forall x \in \mathbb{R}$

$\hookrightarrow$ There are many inequalities that hold for convex and strongly convex inequalities (cheat sheet by Fabian Pedregosa)

Ex) If f convex C^1, then

$$\forall (x,y) \in (\mathbb{R}^d)^2, \quad (\nabla f(x) - \nabla f(y))^T (x-y) \geq 0$$

"co-coercivity"

Proof: $f(y) \geq f(x) + \nabla f(x)^T (y-x)$

and $f(x) \geq f(y) + \nabla f(y)^T (x-y)$

$= f(y) - \nabla f(y)^T (y-x)$

Sum the two inequalities

$$\cancel{f(x)} + \cancel{f(y)} \geq \cancel{f(y)} + \cancel{f(x)} + (\nabla f(x) - \nabla f(y))^T (y-x)$$

$$0 \geq (\nabla f(x) - \nabla f(y))^T (y-x)$$

$$0 \leq (\nabla f(x) - \nabla f(y))^T (x-y)$$

Exercise. Let f be a C^1, μ-strongly convex function in $\mathbb{R}^d$

and let $x^* = \arg\min_x f(x)$ $\left(\{x^*\} = \arg\min_x f(x) \right)$

Goal: $\forall x \in \mathbb{R}^d$, $\|\nabla f(x)\|^2 \geq 2\mu \left(f(x) - \underbrace{f(x^*)}_{\text{sometimes denoted by } \min_{x \in \mathbb{R}^d} f(x)} \right)$

1) Start with

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2} \|y-x\|^2$$ with fixed $x \in \mathbb{R}^d$, $\forall y \in \mathbb{R}^d$

2) Use $\min_y f(y) = f(x^*)$ and that $g : y \mapsto f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2} \|y-x\|^2$ is strongly convex

$$\nabla g(y) = \nabla f(x) + \mu (y-x)$$

$$f(y) \geq g(y) \quad \forall y \in \mathbb{R}^d \implies \underbrace{\min_{y \in \mathbb{R}^d} f(y)}_{= f(x^*)} \geq \min_{y \in \mathbb{R}^d} g(y) \quad (1)$$

g is a strongly convex function (sum of $y \mapsto f(x) + \nabla f(x)^T(y-x)$, that is convex, and $y \mapsto \frac{\mu}{2}\|y-x\|^2$, that is μ-strongly convex)

hence it has a unique global minimum, which is the solution

of $\nabla g(y) = 0 \iff \nabla f(x) + \mu(y-x) = 0$ $\quad \mu > 0$

$$\iff y = x - \frac{1}{\mu}\nabla f(x)$$

Let $y^* := x - \frac{1}{\mu}\nabla f(x)$. Then (1) becomes

$$f(x^*) = \min_{y \in \mathbb{R}^d} f(y) \geq \min_{y \in \mathbb{R}^d} g(y) = g(y^*) \quad (2)$$

$u^T u = \|u\|^2$

$$\begin{aligned} g(y^*) &= f(x) + \nabla f(x)^T(y^* - x) + \frac{\mu}{2}\|y^* - x\|^2 \\ &= f(x) + \nabla f(x)^T\left(-\frac{1}{\mu}\nabla f(x)\right) + \frac{\mu}{2}\left\|-\frac{1}{\mu}\nabla f(x)\right\|^2 \\ &= f(x) - \frac{1}{\mu}\|\nabla f(x)\|^2 + \frac{\mu}{2} \times \frac{1}{\mu^2}\|\nabla f(x)\|^2 \\ &= f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2 \end{aligned}$$

Plugging this expression into (2), we get

$$f(x^*) \geq f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2$$

$$\|\nabla f(x)\|^2 \geq 2\mu\big(\underbrace{f(x) - f(x^*)}_{\geq 0}\big)$$

Takeaway: strongly convex functions are lower bounded by a family of strongly convex quadratic functions $\left(y \mapsto f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}\|y-x\|^2\right)$, and these are very nice functions to optimize

Lipschitz continuous function and optimization

Def: $\varphi : \mathbb{R}^d \to \mathbb{R}^m$ is a L-Lipschitz continuous function if
$\forall (x,y) \in (\mathbb{R}^d)^2, \quad \|\varphi(x) - \varphi(y)\| \leq L \|x-y\| \qquad (L>0)$

Ex) $\varphi : x \mapsto Wx+b$ is $\|W\|$-Lipschitz (continuous)
where $\|W\| = \max_{x \neq 0} \frac{\|Wx\|}{\|x\|}$

NB: L is not unique, but typically the smallest possible value is chosen.

Def: A C^1 function $f : \mathbb{R}^d \to \mathbb{R}$ is called $\left\{ \begin{array}{l} \text{gradient-Lipschitz} \\ C_L^{1,1} \\ L\text{-smooth} \end{array} \right.$ with $L>0$
if $\forall (x,y) \in (\mathbb{R}^d)^2, \quad \|\nabla f(x) - \nabla f(y)\| \leq L \|x-y\|$,
i.e. $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz

Ex) . Any quadratic function $x \mapsto c + g^T x + \frac{1}{2} x^T H x$
$c \in \mathbb{R}, \ g \in \mathbb{R}^d, \ H \in \mathbb{R}^{d \times d}$
is $C_L^{1,1}$ with $L = \left\| \frac{H + H^T}{2} \right\|$

- In particular, $x \mapsto \frac{1}{2n} \|Ax - y\|^2 = \frac{1}{2n} y^T y - \frac{1}{n} y^T A x + \frac{1}{2n} x^T A^T A x$
is $C_L^{1,1}$ with $L = \frac{\|A^T A\|}{n}$
- $x \mapsto \frac{1}{n} \sum_{i=1}^{n} \log(1 + e^{-y_i a_i^T x})$ is $C_L^{1,1}$, $L = \frac{\|A^T A\|}{4n}$

Theorem: Let $f : \mathbb{R}^d \to \mathbb{R}$ is $C_L^{1,1}$, $L>0$ (⚠ not necessarily convex)

Then, $\forall (x,y) \in (\mathbb{R}^d)^2$,

$$f(y) \leq f(x) + \nabla f(x)^T (y-x) + \frac{L}{2} \|y-x\|^2$$

Additional properties

$\to$ If $f \in C_L^{2,1} (= C^2 \cap C_L^{1,1})$, $\|\nabla^2 f(x)\| \leq L$
$\nabla^2 f(x) \preceq L I$

$\to$ If f is $C_L^{1,1}$ and μ-strongly convex,
then $\forall (x,y) \in (\mathbb{R}^d)^2$,

$$\underbrace{f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2} \|y-x\|^2}_{\mu\text{-strong convexity}} \leq f(y) \leq \underbrace{f(x) + \nabla f(x)^T (y-x) + \frac{L}{2} \|y-x\|^2}_{L\text{-smoothness}}$$

Hence $\mu \leq L$
and f can be "sandwiched" between
two (simple) quadratics (aka quadratic functions)

Summary:
- Convex functions are good (for optimization)
  - $\to$ Nice optimality conditions for global optimality
  - $\to$ Even nicer results for strongly convex functions
  - $\to$ Useful inequalities

True for strongly convex quadratic functions $\leftarrow$ [• Best case: $C_L^{1,1}$ + μ-strongly convex function

Last-minute inequality:

f μ-strongly convex and C^1, then

$$\forall x \in \mathbb{R}^d, \quad f(x) - f(x^*) \geq \frac{\mu}{2} \|x - x^*\|^2$$