# 卒研ゼミ「深層学習」



2019年12月9日？

岡崎健人

## 6.2.1.1 Learning Conditional Distributions with Maximum Likelihood

現代におけるNNのほとんどは最尤推定を用いている

$$J(\boldsymbol{\theta}) = -\mathbb{E}_{\mathbf{x},\mathbf{y}\sim\hat{p}_{\text{data}}}\left[\log p_{\text{model}}(\boldsymbol{y}\,|\,\boldsymbol{x})\right]$$

- 最尤推定では、コスト関数は「負の対数尤度」になる
- 訓練データとモデル分布の交差エントロピーと考えることもできる

# ※復習：最尤推定(§5.5)

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} p_{\mathrm{model}}(\mathbb{X};\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{x}\sim\hat{p}_{\mathrm{data}}}\left[\log p_{\mathrm{model}}(\boldsymbol{x};\boldsymbol{\theta})\right]$$

$p_{\mathrm{data}}(\mathbf{x})$：真であるが $\boldsymbol{\theta}$ がわからないため未知の分布

$p_{\mathrm{model}}(\boldsymbol{x};\boldsymbol{\theta})$：値が $\boldsymbol{x}$ であるデータを得る確率（$\boldsymbol{\theta}$ が変数の関数）

$\mathbb{X} = \left\{\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(m)}\right\}$：i. i. d. に従うデータ

$\mathbb{E}_{\mathbf{x}\sim\hat{p}_{\mathrm{data}}}[\,\cdot\,]$：確率変数 $\mathbf{x}$ が経験的な分布 $\hat{p}_{\mathrm{data}}$に従うときの期待値

# ※復習：交差エントロピー(§3.13)

$$
\begin{aligned}
H\left(\hat{p}_{\mathrm{data}}, p_{\mathrm{model}}\right) &= H\left(\hat{p}_{\mathrm{data}}\right) + D_{\mathrm{KL}}\left(\hat{p}_{\mathrm{data}} \| p_{\mathrm{model}}\right) \\
&= -\mathbb{E}_{\mathbf{x} \sim \hat{p}_{\mathrm{data}}}\left[\log p_{\mathrm{model}}(\boldsymbol{x})\right]
\end{aligned}
$$

**注意！**　ここでは訓練データをまとめて１つの文字 $\boldsymbol{x}$ で表しているが、
教科書の該当範囲では $\boldsymbol{x}$ と $\boldsymbol{y}$ が訓練データ

## 6.2.1.1 Learning Conditional Distributions with Maximum Likelihood

**計算例：** $J(\boldsymbol{\theta}) = \dfrac{1}{2}\mathbb{E}_{\mathbf{x},\mathbf{y}\sim\hat{p}_{\text{data}}}\|\boldsymbol{y} - f(\boldsymbol{x};\boldsymbol{\theta})\|^2 + \text{const}\,.$

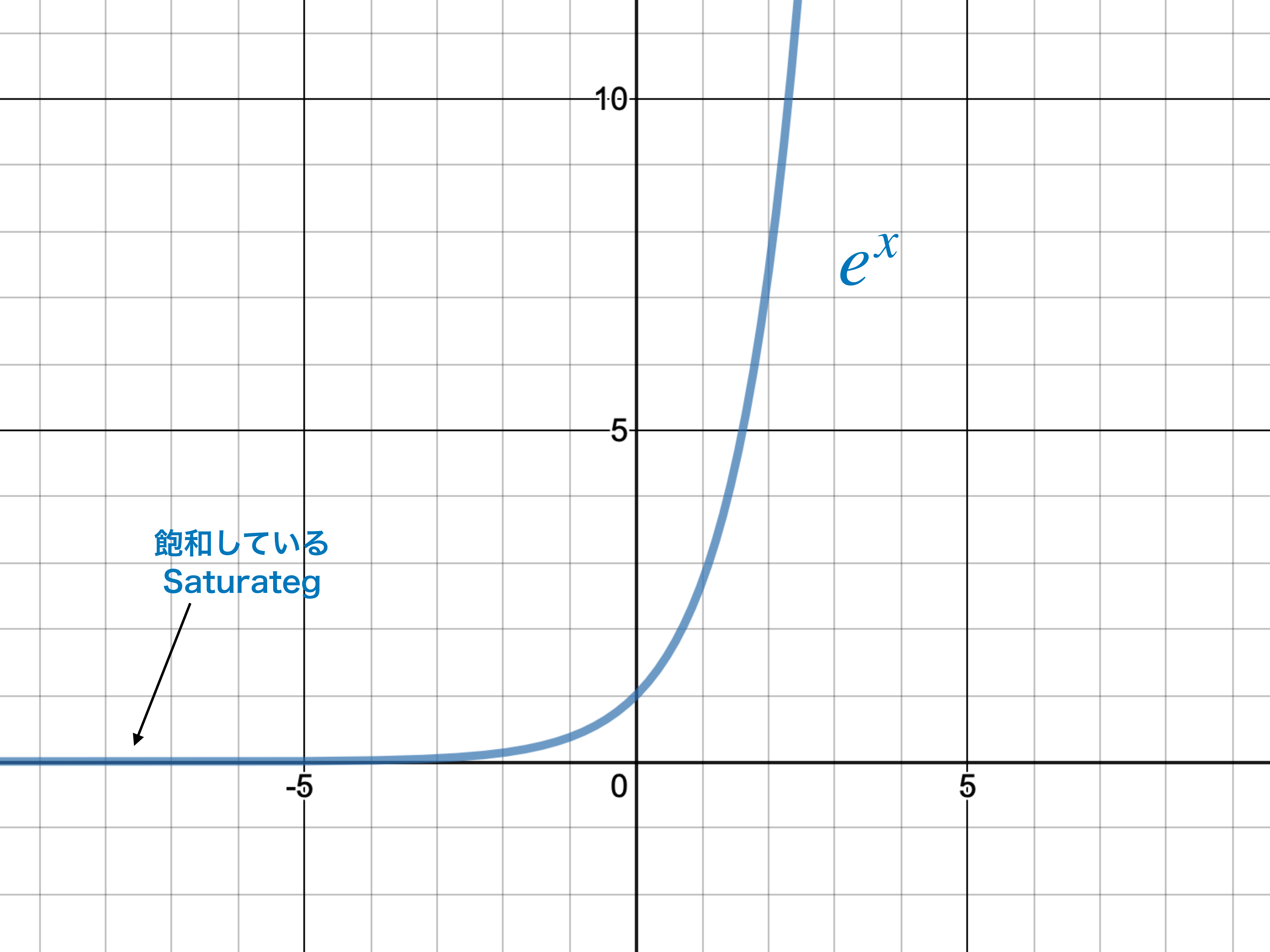
- 尤度関数が$\mathcal{N}(\boldsymbol{y}; f(\boldsymbol{x};\boldsymbol{\theta}), \boldsymbol{I})$に従う場合
- 線形モデルのとき、最尤推定と最小二乗法は等価（§5.5.1）

**コスト関数を自動的に作り出せるのが、最尤推定の強み**

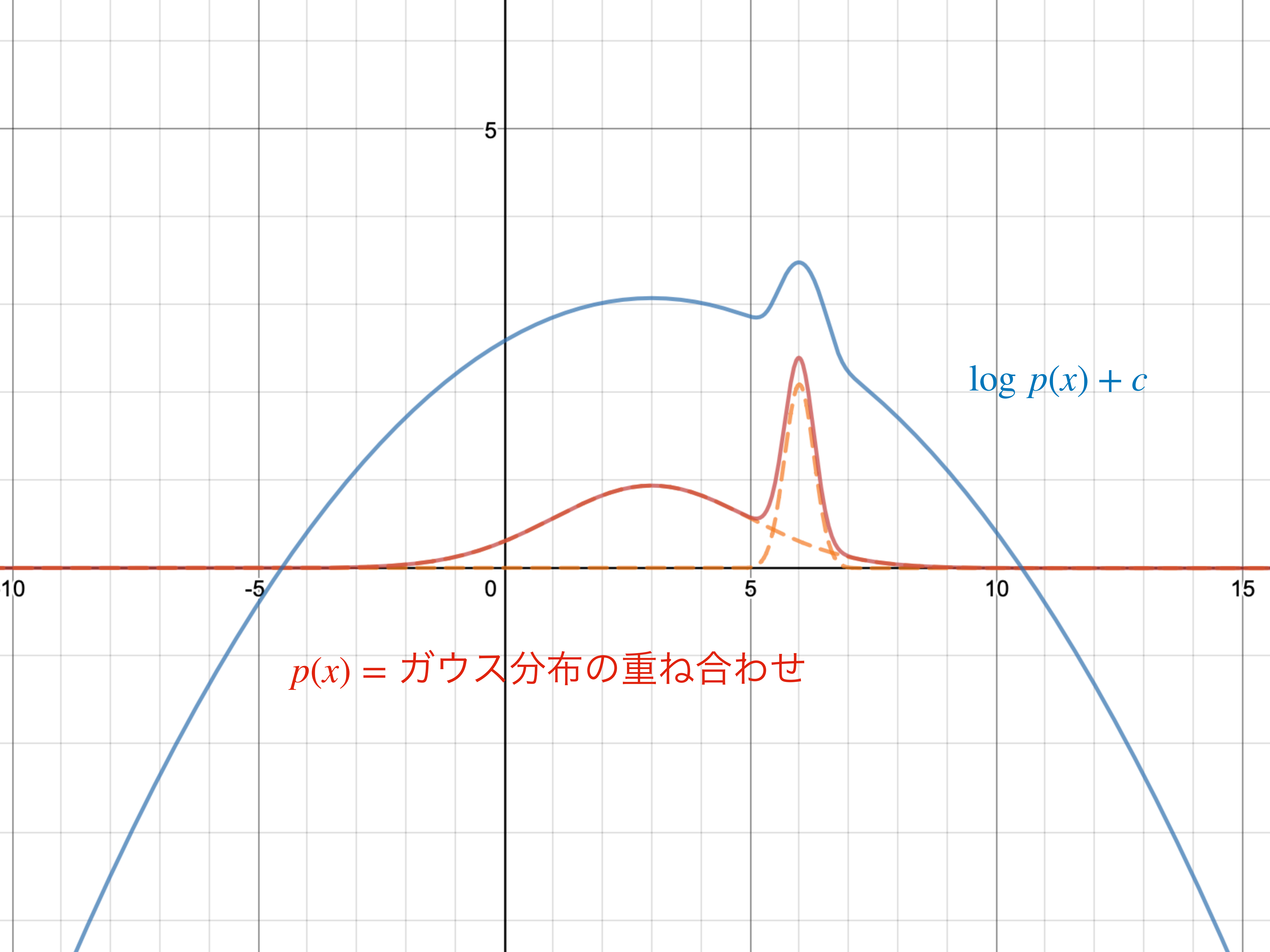
## 6.2.1.1 Learning Conditional Distributions with Maximum Likelihood

### コスト関数の勾配 $|\nabla J(\boldsymbol{\theta})|$ は大きく予測可能であってほしい

- 勾配降下法では $\nabla J(\boldsymbol{\theta})$ を用いて最小点を求めるため
- **飽和する**(saturate, become very flat)ような関数は使えない
- 活性化関数が飽和するとよくない
- 尤度関数が $e^x$ を含んでいても、負の対数尤度は「飽和」しない



10
5
-5
0
5
$e^x$
飽和している
Saturateg



5
-10
-5
0
5
10
15
log $p(x) + c$
$p(x)$ = ガウス分布の重ね合わせ

# 6.2.1.1 Learning Conditional Distributions with Maximum Likelihood

- 実際によく用いられるモデル(？)に対して、$J(\boldsymbol{\theta}) = -\mathbb{E}_{\mathbf{x},\mathbf{y}\sim\hat{p}_{\text{data}}}\left[\log p_{\text{model}}(\boldsymbol{y} \mid \boldsymbol{x})\right]$ は、通常、最小値をもたない
- 連続型確率変数の出力分布の密度を極めて高くできる
- →交差エントロピーは$-\infty$に近づく
- 無限大の報酬を回避するように、正則化を行う（§7）

# 6.2.1.2 Learning Conditional Statistics

## 分布$p(y \mid x; \theta)$よりも「$x$が与えられた元での$y$の推定量」を知りたいとき

- $\boldsymbol{y}$の平均を与える関数 $f(\boldsymbol{x}; \boldsymbol{\theta})$ は知っているものとする
- 十分強力なNNを用いれば、どんな関数 $f$ も表現できる
- この考えは**汎関数**(functional, 関数から実数値へのマッピング)になる
- コスト関数が「ある関数形」をとれば、交差エントロピーが最小値を持つようになる（かも）
- **変分法**(calculus of variations)が必要だが**次のことを知っとけばOK**

# 6.2.1.2 Learning Conditional Statistics

## 知っとくこと①/2

$$f^* = \arg\min_f \mathbb{E}_{\mathbf{x},\mathbf{y}\sim p_{\text{data}}} \|\boldsymbol{y} - f(\boldsymbol{x})\|^2$$

$$\rightarrow \quad f^*(\boldsymbol{x}) = \mathbb{E}_{\mathbf{y}\sim p_{\text{data}}(\boldsymbol{y}|\boldsymbol{x})}[\boldsymbol{y}]$$

無限に多くのデータを真の分布から得るならば、コスト関数は平均二乗誤差になり、$\boldsymbol{x}$ の値が与えられれば対する $\boldsymbol{y}$ 平均を知ることができる

# 6.2.1.2 Learning Conditional Statistics

## 知っとくこと②/2

$$f^* = \arg\min_f \mathbb{E}_{\mathbf{x},\mathbf{y}\sim p_{\text{data}}} \|\boldsymbol{y} - f(\boldsymbol{x})\|_1$$

$$\rightarrow \quad f^*(\boldsymbol{x}) = (\boldsymbol{x}\text{ に対する }\boldsymbol{y}\text{ の中央値})$$

- $\boldsymbol{x}$ の値が与えられたときの $\boldsymbol{y}$ の**中央値**(median)が得られる
- このコスト関数は**平均絶対誤差**(mean absolute error)とよばれる

## 6.2.1.2 Learning Conditional Statistics

### それでも$p(\mathbf{y} \mid \mathbf{x})$を求めたほうがいい

- これらのコスト関数を勾配法で使うと、出力ユニットが飽和してうまくいかないことがある
- そのため条件付き分布$p(\mathbf{y} \mid \mathbf{x})$を求める必要がなくても、交差エントロピーの方法を使うほうがポピュラー