Deep Learning 輪読会 2017

第7章　深層学習のための正則化

2017.11.20

東京大学 大学院 情報理工学系研究科 システム情報学専攻
中村・近藤研究室
東 耕平

構成

深層学習のための正則化

7.1 パラメータノルムペナルティ

7.2 条件付き最適化としてのノルムペナルティ

7.3 正則化と制約不足問題

7.4 データ集合の拡張

7.5 ノイズに対する頑健性

7.6 半教師あり学習

7.7 マルチタスク学習

7.8 早期終了

7.9 パラメータ拘束とパラメータ共有

7.10 スパース表現

7.11バギングやその他のアンサンブル手法

7.12 ドロップアウト

7.13 敵対的学習

7.14 接距離、接伝播法、そして多様体接分類器

深層学習のための正則化

- 正則化
  - 訓練誤差ではなく、汎化誤差の削減を意図した、学習アルゴリズムに対するあらゆる改良
    - パラメータの値を制限
    - パラメータの制限に対応する項を目的関数に加える
    - etc …

- 深層学習の観点から
  - バイアスの増加とバリアンスの減少の引き換えによる推定量の正則化
  - 深層学習の場合、モデルのサイズだけでモデルの複雑さの制御は不可
  - 実践的には、適切に正則化されたモデルが良い性能を発揮

7.1 パラメータノルムペナルティ

- 正則化のアプローチの多くは以下の目的関数 $\tilde{J}$ を最小化

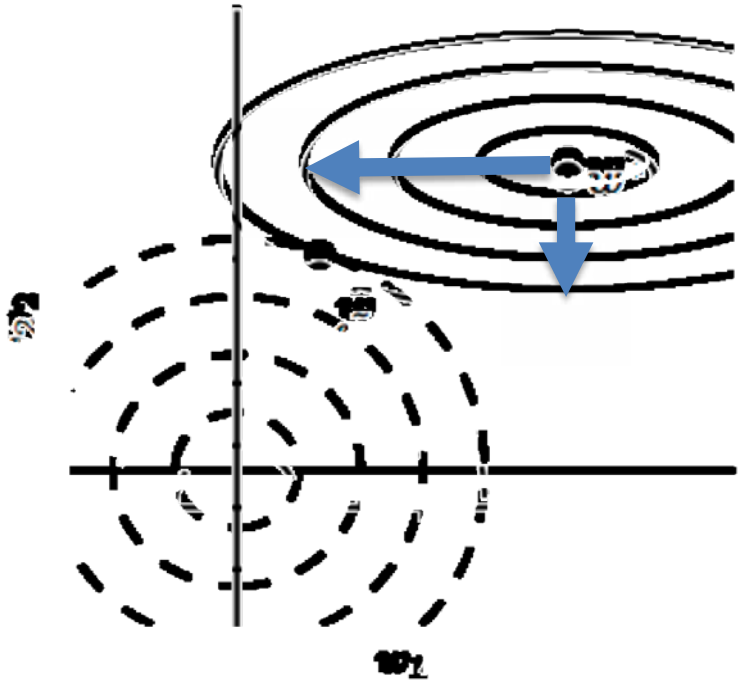
$$\tilde{J}(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}) = J(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}) + \alpha\Omega(\boldsymbol{\theta})$$

J ：目的関数
$\Omega(\boldsymbol{\theta})$：パラメータノルムペナルティ
α：ハイパーパラメータ

- ニューラルネットワークの場合はアフィン変換の重みのみ正則化
  - NNでは $\boldsymbol{\theta}$ はアフィン変換の重み $\boldsymbol{w}$ とバイアス項
  - バイアス項の方が重みより寄与が小さいため
    - バイアスまで正則化すると過小適合の可能性

7.1.1 L^2 パラメータ正則化

- 重み減衰の項を目的関数に追加
  - $\Omega(\boldsymbol{\theta}) = \frac{\mathbf{1}}{\mathbf{2}} ||\boldsymbol{w}||_{\mathbf{2}}^{\mathbf{2}}$
  - 重みを原点に近づける効果
- 正則化の効果を表す図
  - 楕円：正則化されていない目的関数の値が等しい点を結んだ曲線

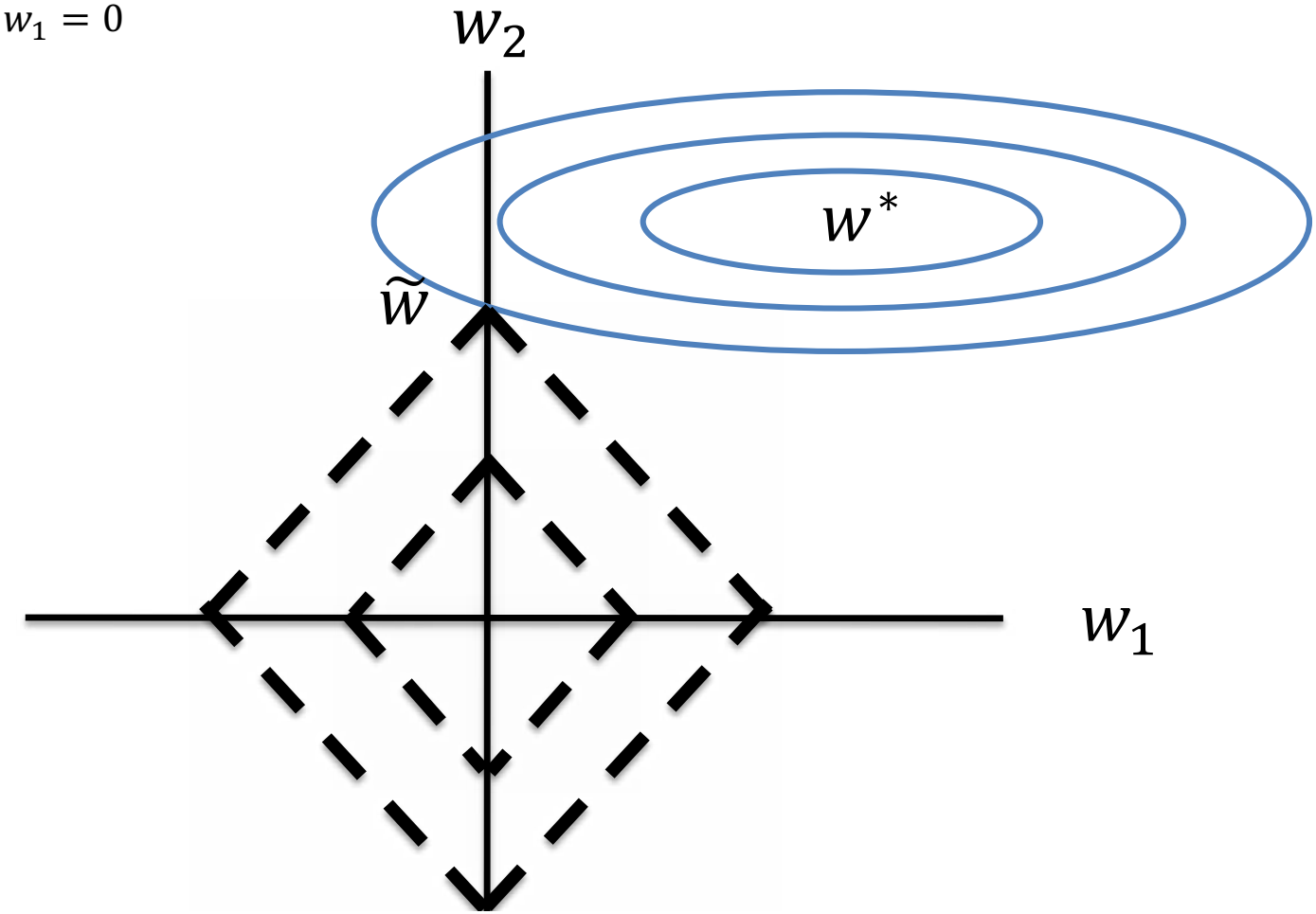


水平方向の移動に対して目的関数の変化は小

垂直方向の移動に対して目的関数の変化は大

7.1.2 L^1 正則化

- パラメータの絶対値の総和を目的関数に追加
  - $\Omega(\boldsymbol{\theta}) = ||\boldsymbol{w}||_1 = \Sigma_i |w_i|$
  - スパースな解が得られる効果 (特徴量選択)
  - 図の場合 $w_1 = 0$



7.2 条件付き最適化としてのノルムペナルティ

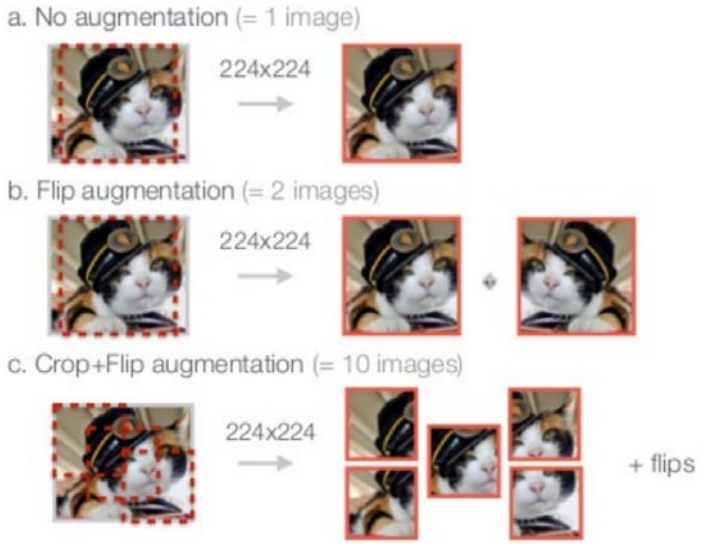
- パラメータノルムペナルティは重みに制約を課す項
  - $\Omega(\boldsymbol{\theta})$ に制約をつけたラグランジュ関数を考え α を固定すると $\tilde{J}$ の最小化の式と等価
  - L2の場合は球内、L1の場合はひし形内
  - α 大：領域 小, α 小：領域 大
- 明示的な制約と再射影を用いる場合
  - Ex. 確率勾配でパラメータを学習した後、 $\Omega(\boldsymbol{\theta}) < \boldsymbol{k}$ を満たす領域に射影
  - 利点
    - 適切なkの値が既知でkに対応するαの探索に時間を使いたくない場合に有用
    - 局所最適解を回避
    - 最適化の安定化

7.3 正則化と制約不足問題

- 問題の適切な定義のため、正則化が必要な場合がある
  - 線形回帰やPCAでは $\boldsymbol{X^T X}$ が特異だと解けない
  - $\boldsymbol{X^T X + \alpha I}$ に置き換えて解く（可逆であることが保証されている）
- 問題が劣決定となってしまう場合
  - 劣決定：解が一意に定まらない
    - ある***w***で完璧な分類を行えるとき、***2w***でも完璧な分類が行えかつ尤度が高くなる
  - 確率勾配降下法の場合、オーバーフローが発生
    - 重み減衰の場合は、尤度の傾き = 重み減衰の傾き となった場合に学習集を終了

7.4 データ集合の拡張

- 元のデータを加工し、新たな訓練データとして学習
  - 一般により多くのデータで学習することで汎化性能は向上
  - 適用できるタスク範囲が限定
    - 密度推定問題では不可
    - 物体認識で特に効果を発揮
    - 音声認識でも効果を発揮
  - ベンチマークテストを行うときは、データ拡張の効果を考慮



図： http://imatge-upc.github.io/telecombcn-2016-dlcv/slides/D2L2-augmentation.pdf

7.5 ノイズに対する頑健性

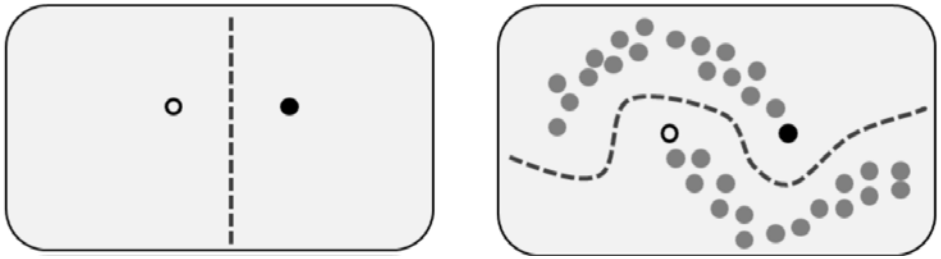
- モデルによってはノイズの入力への追加は重みノルムにペナルティを課すことと等価
  - 一般的には、ノイズの追加(特に隠れユニットに追加する場合)は、単純にパラメータを縮小するより強力
- 重みにノイズを追加する場合
  - モデルの重みは不確実性を持つというベイズ的観点
  - いくつかの仮定の下でノイズの重みへの追加は従来の正則化の形式と等価

7.5.1 出力目標へのノイズの注入

- ラベルのノイズをモデル化
  - データ集合でyのラベルが全て正しいとは限らない
  - ラベル平滑化
    - k個分類を行うソフトマックスの出力値を 0, 1 から $\frac{\epsilon}{k-1}, 1-\epsilon$ へ
    - 重み減衰と組み合わせて学習を安定化

7.6 半教師あり学習

- 教師ありデータに教師なしデータも加えて学習
  - 教師ありデータ: 作成コスト 大　教師なしデータ： 作成コスト　小
  - 1. ブートストラップ法
    - 教師ありデータで学習した分類器で教師なしデータにラベルを付け、そのうち確信度の高い分類結果を教師ありデータに追加して再学習
  - 2. 表現方法の獲得
    - 右は教師ありのみ。左は教師なしで2つの集団を分けてから教師ありで分類。

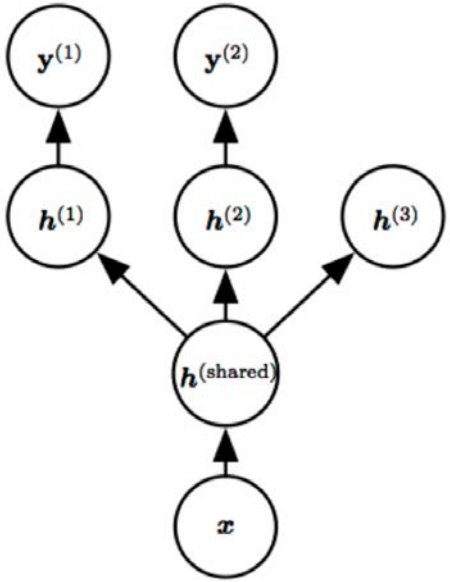


- 深層学習の観点から
  - 目標：表現 $\boldsymbol{h} = f(\boldsymbol{x})$ の学習
    - 同じクラスの事例が類似の表現を持つように表現を学習
    - *Ex) VAE, Ladder Network …*

図：https://commons.wikimedia.org/wiki/File:Example_of_unlabeled_data_in_semisupervised_learning.png

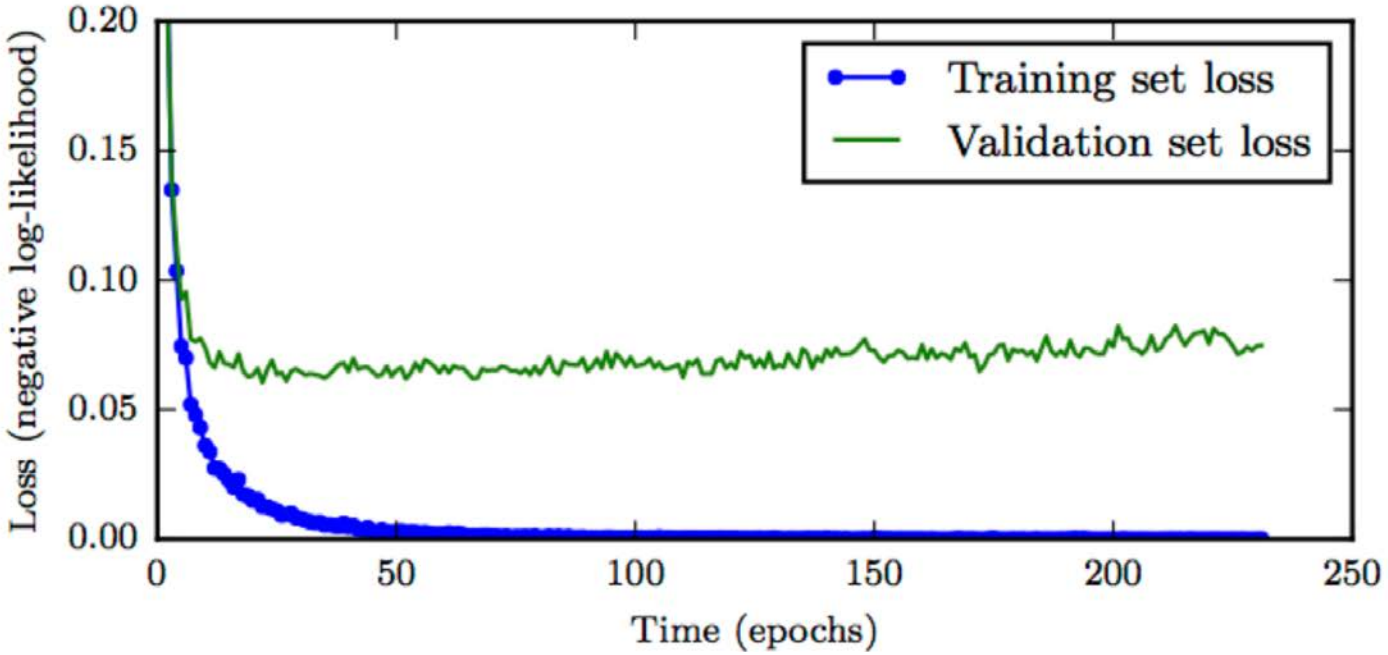
7.7 マルチタスク学習

- いくつかの事例を貯めることで汎化性能を改善する手法
  - モデルの一部をタスク間で共有
  - モデルの共有が妥当だと仮定すれば汎化が改善
- 深層学習の観点から
  - 深層学習では図の共有部分が下位層に相当
  - ""異なるタスクに関連づけられているデータで観測される変動を説明する因子の中には、2つ以上のタスクの間で共有されるものがいくつか存在する""



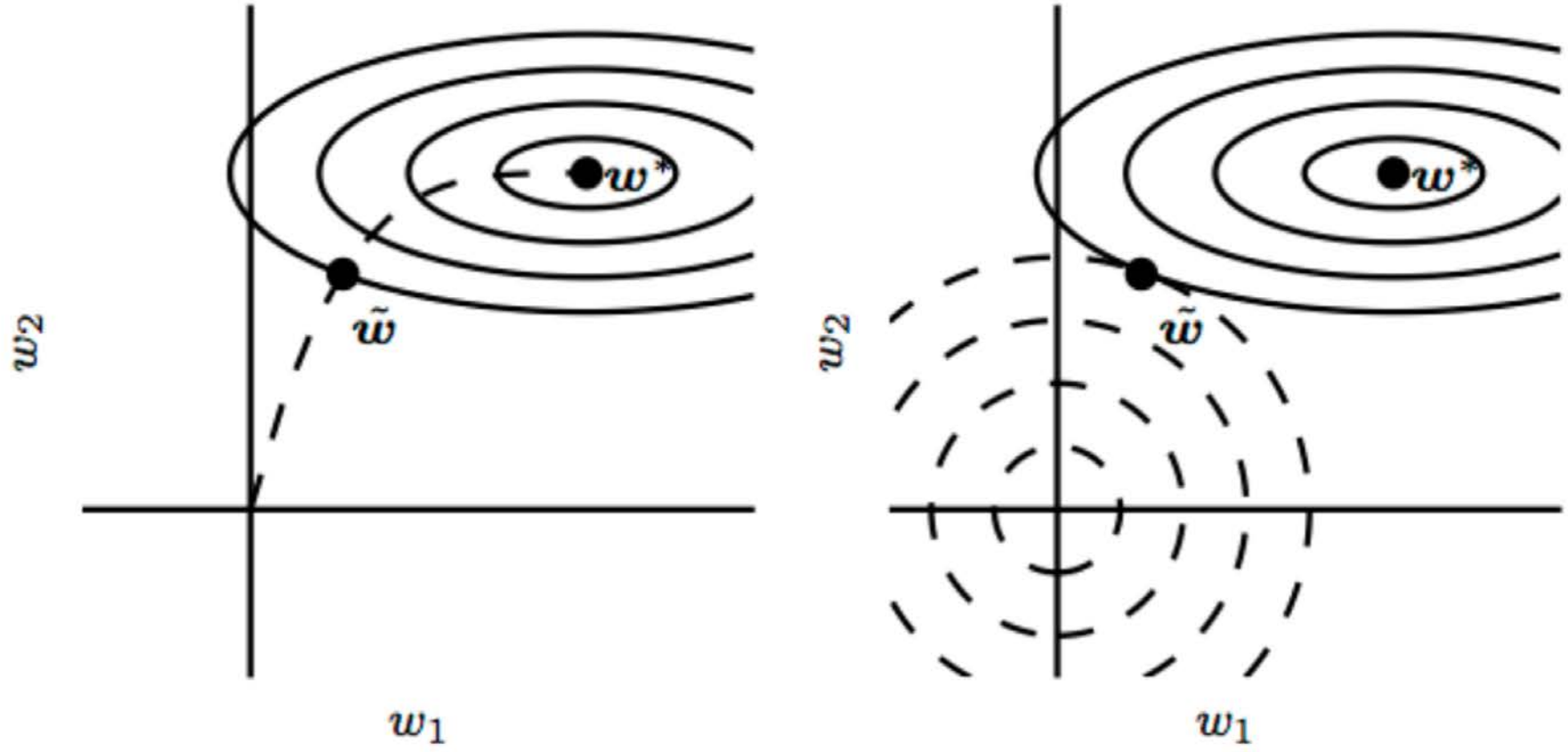
7.8 早期終了

- 適度な学習回数で学習を終了させる正則化手法
  - 訓練誤差は時間とともに減少するが、検証誤差は再び増加してしまう (過剰適合)
  - 訓練の終了回数の決定
    - ハイパーパラメータとして学習回数を事前に決定
    - 検証誤差が連続で増加した場合に中断
  - 検証データも訓練データとして組み込む場合もあり



7.8 早期終了

- 早期終了と正則化の関係



7.9 パラメータ拘束とパラメータ共有

- パラメータ拘束
  - モデルパラメータ間の依存性を反映
    - L^2ノルムを使ったペナルティ　$\Omega\left(w^{(A)}, w^{(B)}\right) = ||w^{(A)} - w^{(B)}||_2^2$
    - 教師ありモデルで学習後、パラメータを教師なし学習のモデルに近づける [Lasserre et al. 2006]
- パラメータ共有
  - パラメータの集合が等しくなるように設定
    - 畳み込みニューラルネットワーク

7.10 スパース表現

- スパースな表現：表現の要素の多くが0
  - 表現 $\boldsymbol{h}$ の正則化

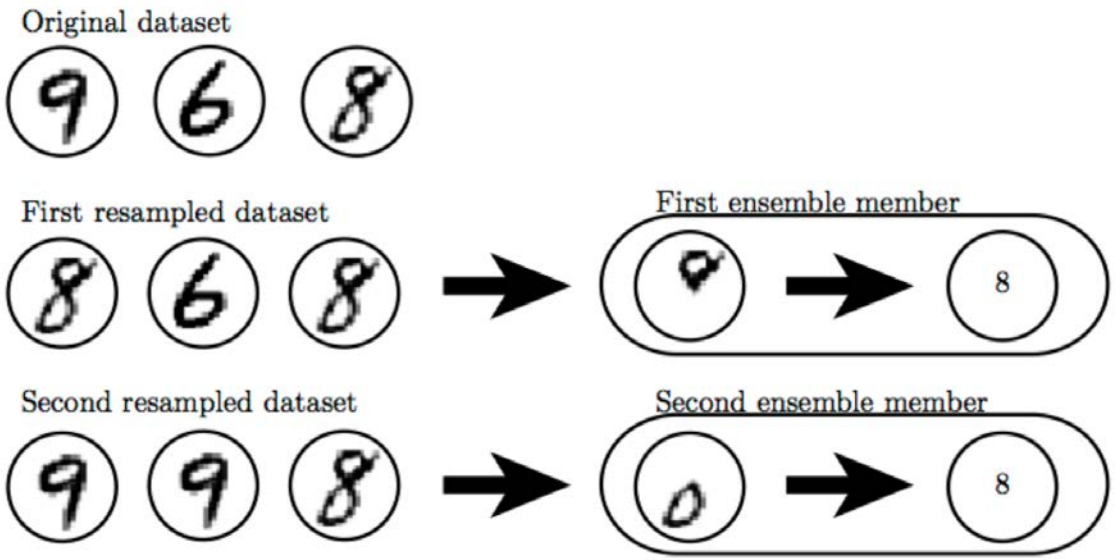
$$\tilde{J}(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}) = J(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}) + \alpha \Omega(\boldsymbol{h})$$

- ユニットの活性化にペナルティを課し、活性化をスパース化
  - 基本的に、隠れユニットを持つモデルはスパース化可
  - 複数の事例の活性化の平均 $\frac{1}{m}\sum_i h^i$ を正則化し、各成分をある値に近づける [Good fellow et al. 2009]
  - 直交マッチング追跡　[Pati et al. 1993]

$$\underset{\boldsymbol{h}, \|\boldsymbol{h}\|_0 < k}{\arg\min} \|\boldsymbol{x} - \boldsymbol{W}\boldsymbol{h}\|^2,$$

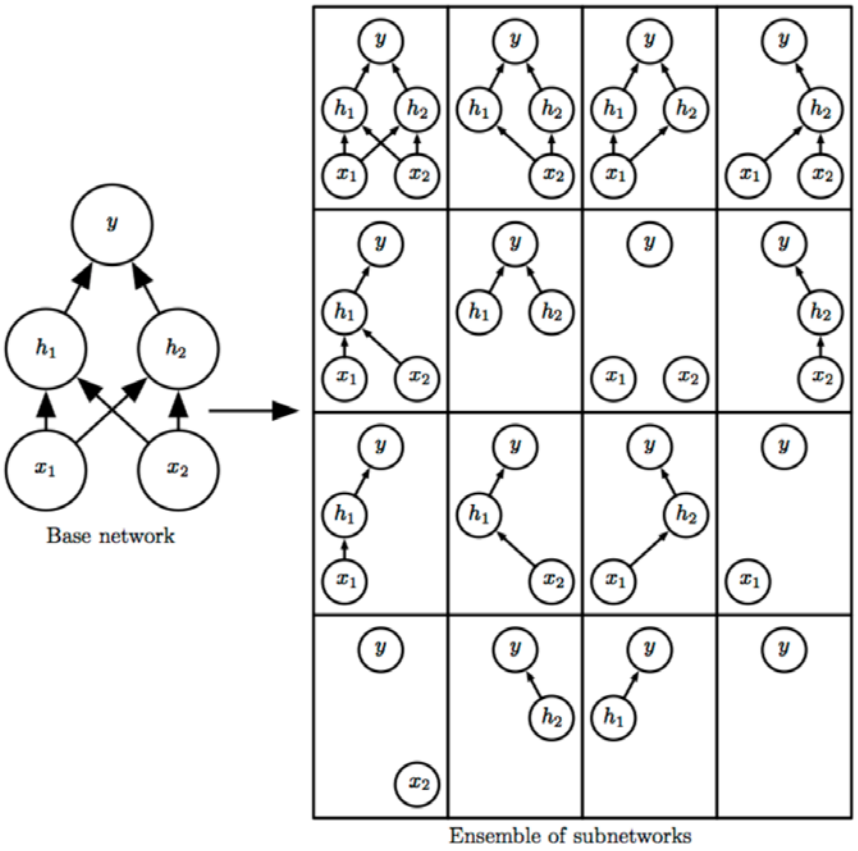
7.11 バギングやその他のアンサンブル手法

- アンサンブル学習
  - モデル平均化：複数のモデルで別々に訓練後、それらの全てのモデルからテストに対しての出力を投票
- バギング
  - アンサンブル学習アルゴリズムの一つ
  - k個の異なるデータ集合の構築
    - 元データ集合から置き換えてサンプリングして構築



7.12 ドロップアウト

- 幅広いモデル族を正則化する、計算量が小さいが強力な正則化手法
  - ミニバッチを入力するたび、すべての入力と隠れユニットを無作為にサンプリング
  - 推論時はすべてのユニットを使用し、各ユニットにユニットがサンプリングされる確率を掛け合わせる（重みスケーリング推論規則）

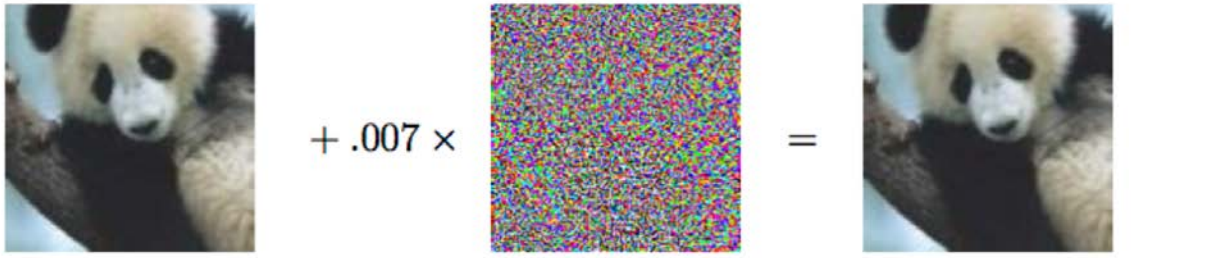


7.12 ドロップアウト

- 利点
  - 計算量が小さい (一個の事例を作る計算量：O(n))
  - 使えるモデルや訓練手続きの種類に重大な制限なし
- ドロップアウトの考察
  - パラメータ共有をするバギングの一形態
    - パラメータ共有、1ステップのみ訓練という点以外はバギングのアルゴリズムに追従
    - ブースティングでは正則化の効果は得られない → ノイズに頑健 < バギングの一種
  - 隠れ層に対する考察
    - モデル間で置換が可能な隠れユニットの獲得
    - 入力ではなく、入力から得られる特徴量を破壊することで、モデルが学習した入力分布への知識をフル活用可

7.13 敵対的学習

- 敵対的学習
  - 訓練集合に敵対的な加工をした事例の学習により精度を向上
- 考察
  - 原因の一つ：過度の線型性
    - 入力の微小変化が出力に大きく影響
  - 敵対的学習はネットワークの訓練データの近傍で局所的に一定にするような学習



7.14 接距離、接伝播法、そして多様体接分類器

- 接距離
  - x_1, x_2 の最近傍距離として、それぞれが属する多様体 M_1, M_2 の間の距離を使用
  - M_i を x_i での接平面で近似して2つの接平面間 or 点と接平面の距離を計算
- 接線伝播法
  - NNの出力*f(x)*を既知の変動要因に対して局所的に不変にするペナルティを追加
    - $\nabla_x f(x)$ が x における接ベクトルに直交 *or* 以下の正則化ペナルティ

$$\Omega(f) = \sum_i \left((\nabla_{\boldsymbol{x}} f(\boldsymbol{x}))^\top \boldsymbol{v}^{(i)} \right)^2 .$$

- 多様体分類器
  - (1) 自己符号化器をつかて教師なし学習で多様体の構造を学習
  - (2) 接線を使って接線伝播法のNN分類器を正則化

参考文献 1

- データ集合の拡張
  - http://imatge-upc.github.io/telecombcn-2016-dlcv/slides/D2L2-augmentation.pdf
- 半教師あり学習
  - http://techon.nikkeibp.co.jp/atcl/mag/15/00144/00009/
  - https://commons.wikimedia.org/wiki/File:Example_of_unlabeled_data_in_semisupervised_learning.png

参考文献 2

- Deep Learning
  - Ian Goodfellow, Yoshua Bengio, Aaron Courville
  - 日本語版
    https://www.amazon.co.jp/%E6%B7%B1%E5%B1%A4%E5%AD%A6%E7%BF%92-Ian-Goodfellow/dp/4048930621