機械学習と
データマイニングの基礎

大阪大学 産業科学研究所
原 聡

原担当パートの内容

- 教師あり学習
  - 11/17（金） 回帰と分類
  - 11/24（金） スパース正則化

- 教師なし学習
  - 12/ 1（金） 密度関数の推定
  - 12/ 8（金） 確率的生成モデル

- 成績評価
  - レポート課題1回（12/1出題）にて評価

教師あり学習

スパース正則化

大阪大学 産業科学研究所
原 聡

教師あり学習の流れ

- データの用意
  - 入力xと出力yを決める。
  - 学習データ$D = \left\{x^{(n)}, y^{(n)}\right\}_{n=1}^{N}$を集める。
- モデルの学習
  - モデル候補（関数のクラス）を設定する。
  - 損失関数、正則化を設定する。
  - データにもっとも適合する関数をモデル候補の中から見つける。
  - 最適なハイパーパラメータを見つける。
- モデルの評価
  - 予測精度の評価

本日の講義内容

応用：特徴選択
Sparse Coding

L1正則化
グループ正則化

最適化
（近接勾配法）

教師あり学習の流れ

- データの用意
  - 入力xと出力yを決める。
  - 学習データ$D = \left\{x^{(n)}, y^{(n)}\right\}_{n=1}^{N}$を集める。
- モデルの学習
  - モデル候補（関数のクラス）を設定する。
  - 損失関数、正則化を設定する。
  - データにもっとも適合する関数をモデル候補の中から見つける。
  - 最適なハイパーパラメータを見つける。
- モデルの評価
  - 予測精度の評価

本日の講義内容

応用：特徴選択
Sparse Coding

L1正則化
グループ正則化

最適化
（近接勾配法）

【例】文章トピックの予測（分類問題）

- 分類：予測したい出力がカテゴリ値の場合
  - 入力：ニュース記事
  - 出力：ニュース記事のトピック（家電 or IT）

トピック： 家電　https://news.livedoor.com/article/detail/5774093/

【ニュース】電力使用量9日が8社管内で今夏最高気温が高い日が続く。特に9日は各地で気温が上がった。気温が上がるとどうしても比例するのが電力使用量だ。9日は、全国の電力会社のうち8社の管内で、いずれも最大電力使用量が今夏最高を記録した。北海道、沖縄電力を除くすべての管内で、この夏一番の電力使用量を記録したが、これはやはり冷房の使用が原因だという。10日も暑くなることが予測されているため、さらに更新する可能性もある。東京電力管内では午後2時台に4824万キロ・ワットを記録。これが今夏の最大使用量を更新したという。電力使用率は88%だった。どうしても気をつけなくてはいけないのが、熱中症だ。電力の使用も気になるが、死亡事故も報告されていることもあり、無理をしないように心がけてもらいたい。最大電力使用量、電力8社管内で今夏最高（読売新聞）

トピック： IT　https://news.livedoor.com/article/detail/6294340/

アップル、デベロッパプレビューをリリース！次期Mac OS X「Mountain Lion」が明らかにアップルは2012年2月16日（米国カリフォルニア州クパティーノ現地時間）、9番目のメジャーリリースとなる「OS X Mountain Lion」のデベロッパプレビューをリリースした。「iPadの人気アプリケーションや機能をMacにもたらし、OS Xのイノベーションを加速させるもの」としているが、どこが凄いのだろうか。MountainLionはメッセージ、メモ、リマインダー、Game Center、通知センター、Share Sheets、Twitterとの統合そしてAirPlayミラーリングをMacに導入するという。また、簡単なセットアップやアプリケーションとの統合ができるようにiCloudが組み込まれた最初のOS Xリリースとなる。デベロッパプレビューには、Gatekeeperも含まれている。これは使用中のMacにどんなアプリケーションがインストールされるかを完全にコントロールすることにより悪意のあるソフトウェアからコンピュータを守る革新的なセキュリティ機能だ。Mountain LionのプレビューリリースはMac Developer Programのメンバーに提供される。アップルの発表では、Macのユーザーは2012年の夏の終わりにMac App StoreからMountain Lionにアップグレードすることができるようになるとしている。■Apple、100以上の新機能を搭載したOS X Mountain Lionのデベロッパプレビューをリリース

【例】文章トピックの予測（分類問題）

- 分類：予測したい出力がカテゴリ値の場合
  - 入力：ニュース記事
  - 出力：ニュース記事のトピック（家電 or IT）

トピック：家電

$y = -1 \qquad x = \left[n_{電}, n_{新}, \ldots\right] = [11, 3, \ldots]$

【ニュース】電力使用量9日が8社管内で今夏最高気温が高い日が続く。特に9日は各地で気温が上がった。気温が上がるとどうしても比例するのが電力使用量だ。9日は、全国の電力会社のうち8社の管内で、いずれも最大電力使用量が今夏最高を記録した。北海道、沖縄電力を除くすべての管内で、この夏一番の電力使用量を記録したが、これはやはり冷房の使用が原因だという。10日も暑くなることが予測されているため、さらに更新する可能性もある。東京電力管内では午後2時台に4824万キロ・ワットを記録。これが今夏の最大使用量を更新したという。電力使用率は88%だった。どうしても気をつけなくてはいけないのが、熱中症だ。電力の使用も気になるが、死亡事故も報告されていることもあり、無理をしないように心がけてもらいたい。最大電力使用量、電力8社管内で今夏最高（読売新聞）

トピック：IT

$y = 1 \qquad x = \left[n_{電}, n_{新}, \ldots\right] = [0, 2, \ldots]$

アップル、デベロッパプレビューをリリース！次期Mac OS X「Mountain Lion」が明らかにアップルは2012年2月16日（米国カリフォルニア州クパティーノ現地時間）、9番目のメジャーリリースとなる「OS X Mountain Lion」のデベロッパプレビューをリリースした。「iPadの人気アプリケーションや機能をMacにもたらし、OS Xのイノベーションを加速させるもの」としているが、どこが凄いのだろうか。MountainLionはメッセージ、メモ、リマインダー、Game Center、通知センター、Share Sheets、Twitterとの統合そしてAirPlayミラーリングをMacに導入するという。また、簡単なセットアップやアプリケーションとの統合ができるようにiCloudが組み込まれた最初のOS Xリリースとなる。デベロッパプレビューには、Gatekeeperも含まれている。これは使用中のMacにどんなアプリケーションがインストールされるかを完全にコントロールすることにより悪意のあるソフトウェアからコンピュータを守る革新的なセキュリティ機能だ。Mountain LionのプレビューリリースはMac Developer Programのメンバーに提供される。アップルの発表では、Macのユーザーは2012年の夏の終わりにMac App StoreからMountain Lionにアップグレードすることができるようになるとしている。■Apple、100以上の新機能を搭載したOS X Mountain Lionのデベロッパプレビューをリリース

【例】文章トピックの予測（分類問題）

- トピック予測データ
  - 家電（$y = -1$）：864件
  - IT（$y = 1$）：870件

  https://www.rondhuit.com/download.html#ldcc より

- 問
  - トピック予測に効果的な漢字は何か？
    - 前回の「京都観光/グルメ」では「京」と「味」でほぼ予測できた。
  - 効果的な漢字がわからない場合
    - とりあえず全部の漢字を候補とする。
      → 全部で1913種類の漢字
    - 出現頻度の低い漢字を間引く。
      → 頻度1%以下の漢字を間引くと、残りは1305種類

問題設定

- 使えるデータが少ない場合のモデルの学習
  - 学習データが大量に手に入らない場合。
    - もしもニュース記事が200件しか手に入らなかったら？

- データセットの分割
  - 学習データ 200件：分類モデルの学習に使用。
    - $\left\{x^{(n)}, y^{(n)}\right\}_{n=1}^{N}$ – $y=-1$：100件、$y=1$：100件、x：1305次元
  - テストデータ 1534件：分類モデルの評価に使用。
    - $\left\{x_{\text{test}}^{(m)}, y_{\text{test}}^{(m)}\right\}_{m=1}^{M}$ – $y=-1$：764件、$y=1$：770件、x：1305次元

【参考】データが少ないと過学習が容易に起こる

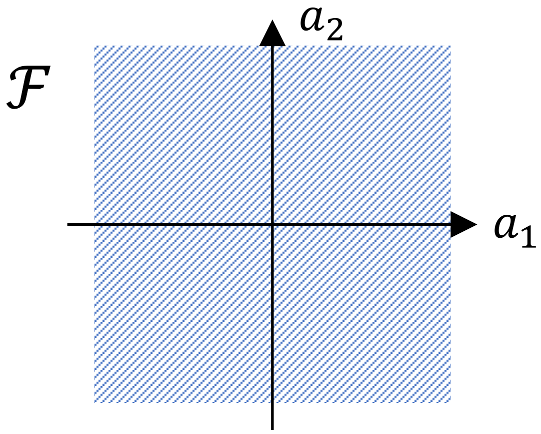
- データの次元dがデータ数Nより大きいならば、
線形分類モデルで学習データを完璧に分類できる。
  - 線形分類モデル
    - $f_\theta(x) = \sum_{i=1}^{d} a_i x_i + b = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \\ 1 \end{bmatrix}^\top \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_d \\ b \end{bmatrix} = x^\top \theta$

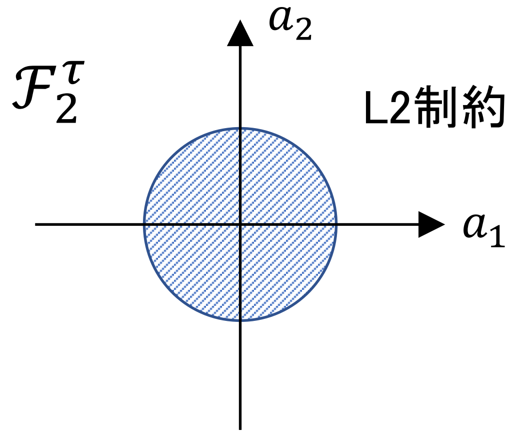
      $x \in \mathbb{R}^{d+1}$　　$\theta \in \mathbb{R}^{d+1}$
  - 完璧な分類 $\Leftrightarrow y^{(n)} {x^{(n)}}^\top \theta > 0, \forall n \in \{1, 2, \dots, N\}$
    - $\epsilon > 0$に対して$Z\theta = \epsilon 1$を満たすθが存在すれば完璧な分類が可能。
      - $Z = \left[y^{(1)} x^{(1)}, y^{(2)} x^{(2)}, \dots, y^{(N)} x^{(N)} \right]^\top \in \mathbb{R}^{N \times (d+1)}$
    - $\mathrm{rank}(Z) = N < d$ならばこのようなθは常に存在する。
      - 独立な式の数より変数の数が多い連立方程式の解は常に存在する。

過学習を抑制する方法：L2正則化

- 過学習を抑制するために関数のクラスを小さくする。

$$\mathcal{F}_2^\tau = \left\{\sum_{i=1}^{d} a_i x_i + b \mid \|a\|^2 \le \tau\right\}$$





- L2正則化付きの経験損失最小化

$$\min_\theta \frac{1}{N}\sum_{n=1}^{N} \ell\big(y, f_\theta(x)\big) + \frac{\lambda}{2}\|w\|^2$$

過学習を抑制する方法：L2正則化

- ロジスティック損失を用いた場合の比較

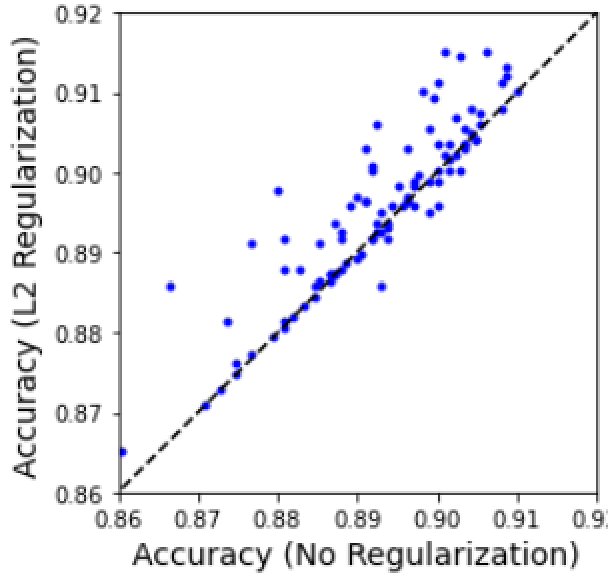
$$\min_{\theta} \frac{1}{N} \sum_{n=1}^{N} \log\left(1 + \exp\left(-y^{(n)} {x^{(n)}}^{\top} \theta\right)\right)$$

正則化なし

$$\min_{\theta} \frac{1}{N} \sum_{n=1}^{N} \log\left(1 + \exp\left(-y^{(n)} {x^{(n)}}^{\top} \theta\right)\right) + \frac{\lambda}{2} \|a\|^2$$

L2正則化

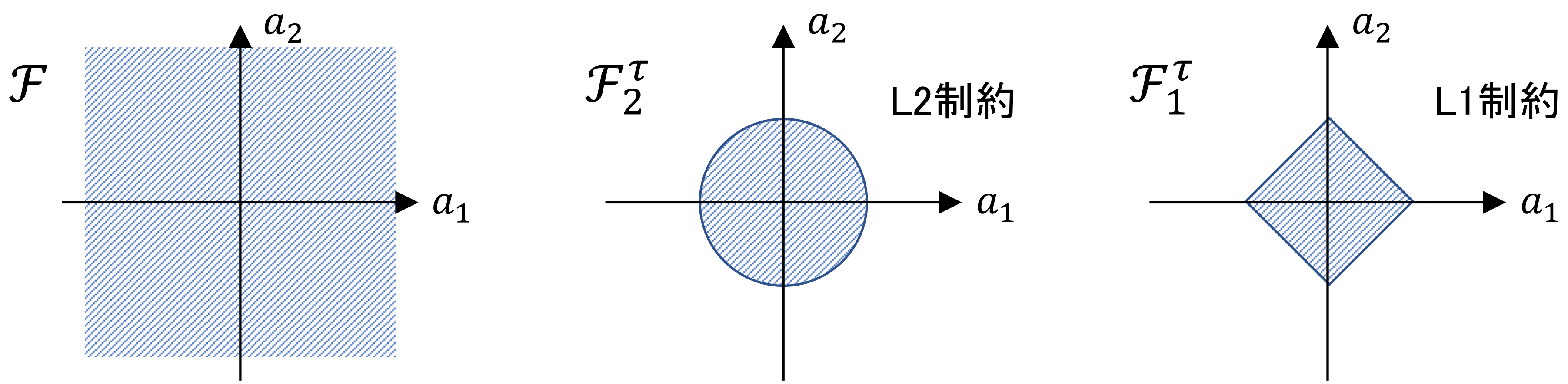
- 実験評価
  - 学習/テストの分割をランダムに100回。
  - テストセットでのAccuracyを評価。
    - λは5-foldの交差検証で決定。
  - L2正則化は過学習の抑制に効果的。
    - L2正則化を使ったほうが、正則化なしよりもAccuracyが高い。



L1正則化

- 過学習を抑制するために関数のクラスを小さくする。

$$\mathcal{F}_1^\tau = \left\{\sum_{i=1}^d a_i x_i + b \mid \sum_{i=1}^d |a_i| =: \|a\|_1 \le \tau\right\}$$

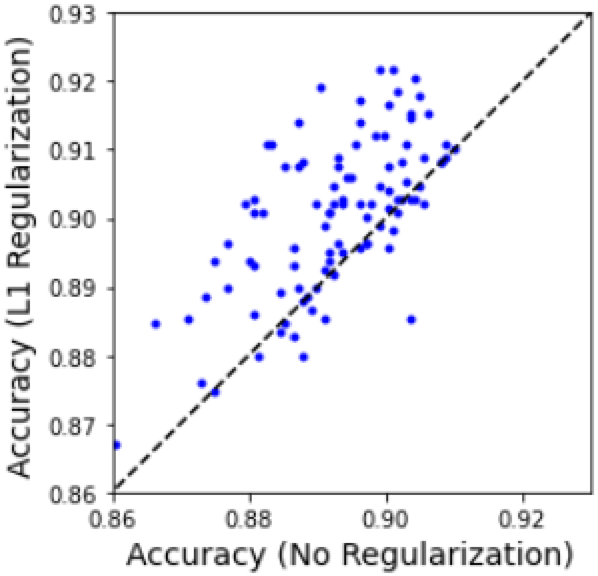


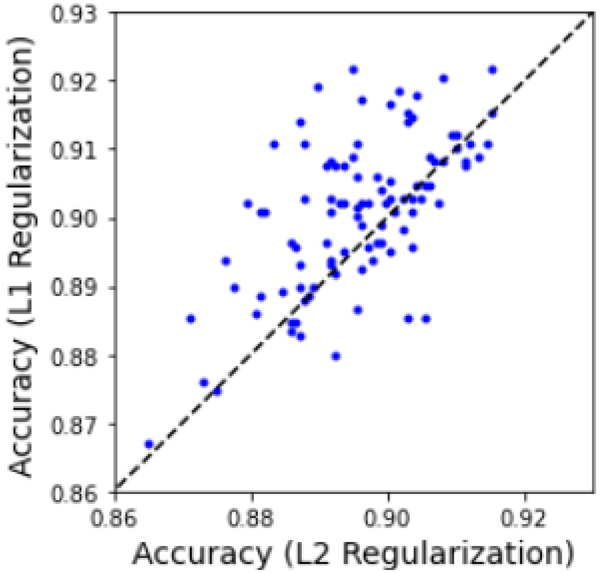
- L1正則化付きの経験損失最小化

$$\min_\theta \frac{1}{N}\sum_{n=1}^N \ell\big(y, f_\theta(x)\big) + \lambda\|a\|_1$$

L1正則化

- データが高次元かつデータ数が少ない場合には、L1正則化のほうがL2正則化よりも効果的なことが多い。





- L1正則化を使うために
  - L1正則化のための最適化法
  - L1正則化の性質の理解

【参考】Lpノルム

- ユークリッドノルム
  - $\|\theta\| = \left(\sum_{i=1}^{d} \theta_i^2\right)^{1/2}$
- Lpノルム（二乗の代わりにp乗を使う）
  - $\|\theta\|_p = \left(\sum_{i=1}^{d} |\theta_i|^p\right)^{1/p}$
    - ユークリッドノルムは$p = 2$の場合に相当: $\|\theta\| = \|\theta\|_2$
    - 無限大ノルム（$p = \infty$）: $\|\theta\|_\infty = \max_i |\theta_i|$
- Hölderの不等式
  - $\alpha^\top \beta \le \|\alpha\|_p \|\beta\|_q \quad \left(\frac{1}{p} + \frac{1}{q} = 1\right)$
    - $p = q = 2$の場合はCauchy–Schwarzの不等式に相当

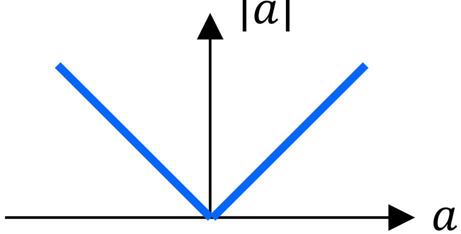
L1正則化を使うために

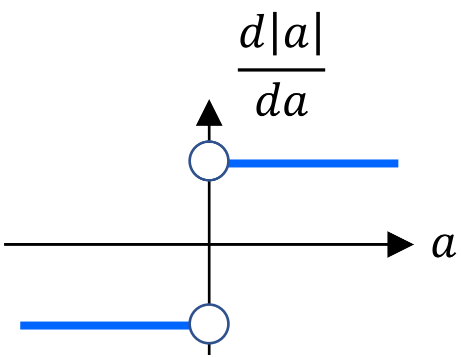
- 1. L1正則化のための最適化法
  - 近接勾配法

- 2. L1正則化の性質の理解
  - スパース性

L1正則化のための最適化法

- L1正則化の最適化の難しさ
  - 原点で微分できない。





- 勾配降下法（最急降下法）は微分が存在しないと使えない。

- 近接勾配法
  - 勾配降下法の拡張版

近接勾配法

- 勾配降下法は目的関数の二次近似を最小化していると解釈できる。

$$\min_{\theta} L(\theta)$$

- 勾配降下法のtステップ目を$\theta^{[t]}$とする。
- $L(\theta)$の二次近似

$$L(\theta) \approx L\left(\theta^{[t]}\right) + \frac{\partial L\left(\theta^{[t]}\right)}{\partial \theta}^{\mathrm{T}} \left(\theta - \theta^{[t]}\right) + \frac{1}{2\eta} \left\|\theta - \theta^{[t]}\right\|^2$$

$$= \frac{1}{2\eta} \left\| \theta - \theta^{[t]} + \eta \frac{\partial L\left(\theta^{[t]}\right)}{\partial \theta} \right\|^2 + L\left(\theta^{[t]}\right)$$

- 近似の最小化

$$\theta = \theta^{[t]} - \eta \frac{\partial L\left(\theta^{[t]}\right)}{\partial \theta}$$

勾配降下法の更新式

近接勾配法

- 二次近似をL1正則化以外の項に適用する。

$$\min_{\theta=[a_1,a_2,\ldots,a_d,b]} L(\theta)+\lambda\|a\|_1$$

- $L(\theta)$の二次近似

$$L(\theta)\approx\frac{1}{2\eta}\left\|\theta-\theta^{[t]}+\eta\frac{\partial L\left(\theta^{[t]}\right)}{\partial\theta}\right\|^2+L\left(\theta^{[t]}\right)$$

- 近似の最小化

$$\min_{\theta}\frac{1}{2}\left\|\theta-\theta^{[t]}+\eta\frac{\partial L\left(\theta^{[t]}\right)}{\partial\theta}\right\|^2+\lambda\eta\|w\|_1$$

この最小化解を$\theta^{[t+1]}$とするのが
近接勾配法

近接勾配法

$$\min_{\theta=[a_1,a_2,\dots,a_d,b]} \frac{1}{2}\left\|\theta-\theta^{[t]}+\eta\frac{\partial L\left(\theta^{[t]}\right)}{\partial\theta}\right\|^2+\lambda\eta\|a\|_1$$

- 一変数の最適化問題への分解

$$\min_b \frac{1}{2}\left(b-b^{[t]}+\eta\frac{\partial L\left(\theta^{[t]}\right)}{\partial b}\right)^2$$

$$b=b^{[t]}-\eta\frac{\partial L\left(\theta^{[t]}\right)}{\partial b}$$

$$\min_{a_i} \frac{1}{2}\left(a_i-a_i^{[t]}+\eta\frac{\partial L\left(\theta^{[t]}\right)}{\partial a_i}\right)^2+\lambda\eta|a_i|$$

- 以下の最適化が解ければ良い。

$$\min_u \frac{1}{2}(u-v)^2+\alpha|u|$$

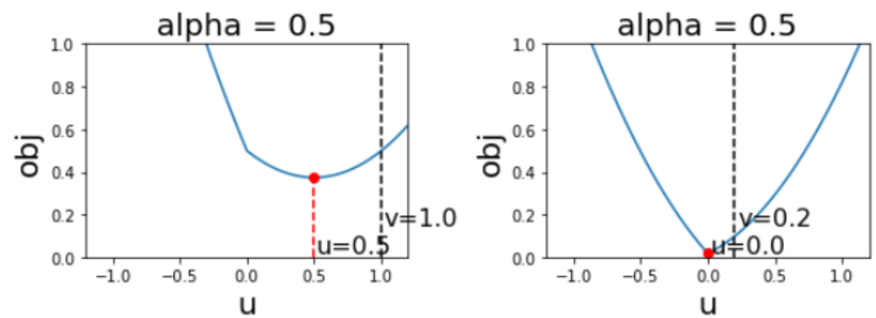
$$\begin{cases} u \leftarrow a_i \\ v \leftarrow a_i^{[t]}-\eta\dfrac{\partial L\left(\theta^{[t]}\right)}{\partial a_i} \\ \alpha \leftarrow \lambda\eta \end{cases}$$

近接勾配法

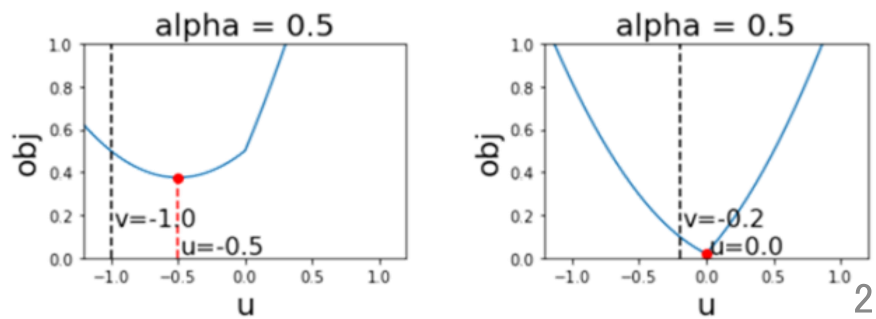
$$\min_u \frac{1}{2}(u-v)^2 + \alpha|u|$$

$u = \mathrm{sign}(v)\max\{0, |v| - \alpha\}$

- 場合分け
  - 1. $v \geq 0$の場合：最適な$u \geq 0$
    - $\min_u \frac{1}{2}(u-v)^2 + \alpha u \quad \rightarrow \quad u = v - \alpha$
    - $v > \alpha$：$u = v - \alpha$が最適
    - $0 \leq v \leq \alpha$：$u = 0$が最適



  - 2. $v \leq 0$の場合：最適な$u \leq 0$
    - $\min_u \frac{1}{2}(u-v)^2 - \alpha u \quad \rightarrow \quad u = v + \alpha$
    - $v < -\alpha$：$u = v + \alpha$が最適
    - $-\alpha \leq v \leq 0$：$u = 0$が最適



近接勾配法

$$\min_{\theta=[a_1,a_2,\ldots,a_d,b]} L(\theta) + \lambda\|a\|_1$$

- 更新式
  - $b^{[t+1]} = b^{[t]} - \eta \frac{\partial L(\theta^{[t]})}{\partial b}$
  - ${a_i}^{[t+1]} = \mathrm{sign}(v_i)\max\{0, |v_i| - \lambda\eta\}$
    - $v_i = {a_i}^{[t]} - \eta \frac{\partial L(\theta^{[t]})}{\partial a_i}$

- 近接勾配法は絶対値の微分を必要としない。
  - $L(\theta)$が微分可能なら問題ない。

【参考】一般の近接勾配法

$$\min_{\theta} L(\theta) + \Omega(\theta)$$

- 問題設定
  - $L(\theta)$は微分可能。$\Omega(\theta)$は微分不可能でもOK。
- 一般の近接勾配法
  - 二次近似の最小化が解ければ良い。

$$\min_{\theta} \frac{1}{2}\|\theta - \beta\|^2 + \eta\Omega(\theta)$$

  - 最小化解を求める操作を $\mathrm{prox}_{\eta\Omega}(\beta)$ と書く。
    - $\mathrm{prox}_{\eta\Omega}(\beta)$を$\eta\Omega$の近接作用素(Proximity Operator)と呼ぶ。
  - 更新式
    - $\theta^{[t+1]} = \mathrm{prox}_{\eta\Omega}\left(\theta^{[t]} - \eta \frac{\partial L(\theta^{[t]})}{\partial\theta}\right)$

近接作用素が効率的に
計算できる問題ならば、
近接勾配法は効率的

【参考】近接勾配法の応用

- 制約付き最適化
  - $\min_{\theta} L(\theta) \;\; \mathrm{s.t.} \;\; g(\theta) \leq 0$
  - $\min_{\theta} L(\theta) + \Omega(\theta)$
    - $\Omega(\theta) = \begin{cases} 0 & \text{if } g(\theta) \leq 0 \\ \infty & \text{if } g(\theta) > 0 \end{cases}$

等価な問題

  - 近接作用素
    - $$\begin{aligned} \mathrm{prox}_{\eta\Omega}(\beta) &= \operatorname*{argmin}_{\theta} \frac{1}{2} \|\theta - \beta\|^2 + \eta\Omega(\theta) \\ &= \operatorname*{argmin}_{\theta} \frac{1}{2} \|\theta - \beta\|^2 \;\; \mathrm{s.t.} \;\; g(\theta) \leq 0 \end{aligned}$$

βを制約条件を満たす空間にユークリッド射影する

  - 例. 非負値制約 $g(\theta) = -\theta$
    - $$\begin{aligned} \mathrm{prox}_{\eta\Omega}(\beta) &= \operatorname*{argmin}_{\theta} \frac{1}{2} \|\theta - \beta\|^2 \;\; \mathrm{s.t.} \;\; \theta \geq 0 \\ &= \max\{0, \beta\} \end{aligned}$$

L1正則化を使うために

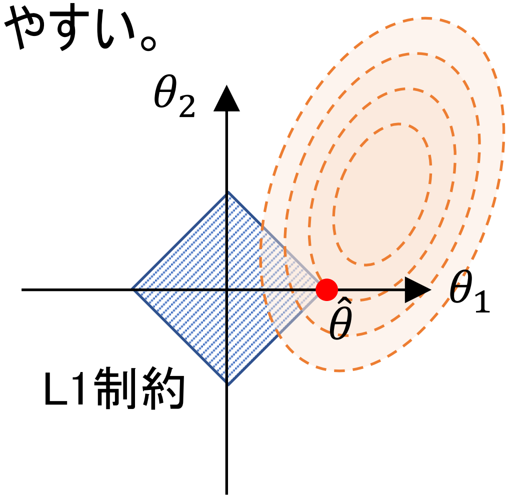
- 1. L1正則化のための最適化法
  - 近接勾配法
- 2. L1正則化の性質の理解
  - スパース性

スパース性

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, L(\theta) + \lambda \|\theta\|_1$$

- $\hat{\theta}$は零要素を持つ傾向（スパース性）がある。
  - $\hat{\theta} = [\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_d]$ のうちいくつかの$\hat{\theta}_i$は$\hat{\theta}_i = 0$となることがある。
  - λが大きい/小さいと$\hat{\theta}_i = 0$となる要素は増える/減る傾向がある。

- なぜスパース性があるか？
  - 制約付き問題の最適解は端点（ゼロ点）に集まりやすい。
    - $\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, L(\theta)$ s.t. $\|\theta\|_1 \le \tau$



  - 更新式はθをゼロに丸め込む。
    - $\theta_i^{[t+1]} = \operatorname{sign}(v_i) \times \max\{0, |v_i| - \lambda\eta\}$
    - $v_i = \theta_i^{[t]} - \eta \dfrac{\partial L(\theta^{[t]})}{\partial \theta_i}$

スパース性に基づく性能保証

- L1制約付きの最小二乗回帰($X, Y \in \mathbb{R}^{N\times d} \times \mathbb{R}^{N}$)を考える。
  - $\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{2}\|X\theta - Y\|^2 \ \text{s.t.}\ \|\theta\|_1 \leq \tau$

- 仮定
  - $Y = X\theta^* + \epsilon$ となる真のθ^*が存在する。
    - θ^*は高々k個の非ゼロ要素を持つ。
  - ϵは平均0, 分散σ^2の正規分布に従う。

- 定理
  - 適当な条件の元で

$$\|\hat{\theta} - \theta^*\| = O\left(\sigma\sqrt{\frac{\mathrm{klogd}}{N}}\right)$$

推定誤差への次元dの影響は$\log d$。
本質的な支配パラメータは非ゼロ要素数 k 。
θ^*がスパースなときにL1制約を使うと
推定誤差が減る。

教師あり学習の流れ

- データの用意
  - 入力xと出力yを決める。
  - 学習データ$D = \{x^{(n)}, y^{(n)}\}_{n=1}^{N}$を集める。
- モデルの学習
  - モデル候補（関数のクラス）を設定する。
  - 損失関数、正則化を設定する。
  - データにもっとも適合する関数をモデル候補の中から見つける。
  - 最適なハイパーパラメータを見つける。
- モデルの評価
  - 予測精度の評価

本日の講義内容

応用：特徴選択
Sparse Coding

L1正則化
グループ正則化

最適化
（近接勾配法）

特徴選択

- 文章トピックの予測
  - トピック予測に効果的な漢字は何か？
    - 頻度1%以下の漢字を間引いても、残りは1305種類
    - 1305種類のうち、どの漢字が予測に効果的か？

- 特徴選択
  - 多数の特徴の中から、予測に有効な特徴を選び出す。
  - スパースな線形モデルの学習 → 特徴選択

$$f_\theta(x) = \sum_{i=1}^{d} a_i x_i + b = \boxed{\sum_{i:a_i \neq 0} a_i x_i + b}$$

（特徴選択）$a_i \neq 0$ となった特徴x_iのみが予測に使われる。

特徴選択

- 文章トピックの予測
  - トピック予測に効果的な漢字は何か？
    - 頻度1%以下の漢字を間引いても、残りは1305種類
    - 1305種類のうち、どの漢字が予測に効果的か？

予測には漢字300個程度で十分

$f_\theta(x)$ = - 0.89 題 - 0.77 筋 - 0.75 話 + 0.46 見 - 0.42 連 + 0.40 永 - 0.40 差 + 0.39 筆 - 0.34 記 + 0.32 何 - 0.28 質 + 0.26 行 + 0.26 如 + 0.26 絵 + 0.26 紺 + 0.24 類 - 0.23 関 + 0.23 最 - 0.23 紀 - 0.23 誰 + 0.23 凄 - 0.22 呂 + 0.21 名 - 0.20 気 + 0.20 時 + 0.19 更 + 0.19 税 - 0.19 影 + 0.18 宝 - 0.18 発 + 0.18 不 - 0.18 夏 + 0.18 速 + 0.18 銅 + 0.18 概 + 0.18 標 + 0.17 定 + 0.17 封 - 0.17 節 + 0.17 跡 - 0.17 牧 + 0.16 遂 + 0.16 考 + 0.16 式 + 0.16 内 + 0.15 英 - 0.15 温 + 0.15 化 + 0.15 起 - 0.15 争 - 0.15 評 + 0.15 剣 - 0.14 被 + 0.14 会 + 0.14 長 - 0.14 順 + 0.14 旧 + 0.14 途 - 0.14 精 + 0.14 変 - 0.13 疲 - 0.13 衝 + 0.13 奏 - 0.13 鍋 + 0.13 理 + 0.13 勝 + 0.13 始 - 0.13 視 + 0.13 募 - 0.12 撃 - 0.12 拓 - 0.12 灯 - 0.12 高 - 0.12 喜 - 0.12 婚 + 0.12 頻 + 0.12 所 + 0.12 則 + 0.12 添 + 0.12 解 + 0.12 荷 + 0.12 使 - 0.12 庭 + 0.11 付 + 0.11 木 + 0.11 間 - 0.11 心 - 0.11 洋 + 0.11 法 + 0.11 介 + 0.11 泳 + 0.11 同 + 0.11 弾 - 0.11 林 - 0.11 暑 + 0.11 十 + 0.10 孫 + 0.10 言 - 0.10 華 + 0.10 紹 - 0.10 史 - 0.10 授 - 0.10 交 - 0.10 炎 - 0.10 挑 - 0.10 浦 + 0.09 遅 + 0.09 復 + 0.09 総 + 0.09 便 - 0.09 現 + 0.09 聞 + 0.09 社 + 0.09 説 + 0.09 償 - 0.09 工 + 0.09 認 + 0.08 択 + 0.08 滴 + 0.08 海 - 0.08 特 - 0.08 市 - 0.08 貯 + 0.08 方 - 0.08 働 + 0.08 権 + 0.08 入 + 0.08 航 - 0.08 冬 + 0.08 完 + 0.08 描 + 0.08 賛 + 0.08 厚 + 0.08 子 - 0.08 恋 + 0.07 岸 + 0.07 様 + 0.07 融 + 0.07 橋 + 0.07 象 + 0.07 触 + 0.07 芸 + 0.07 義 + 0.07 続 + 0.07 曲 - 0.07 験 + 0.07 部 - 0.07 積 - 0.07 襲 + 0.07 利 + 0.07 知 + 0.06 仲 - 0.06 女 + 0.06 余 + 0.06 議 + 0.06 果 - 0.06 去 - 0.06 禁 - 0.06 丈 - 0.06 撤 - 0.06 因 + 0.06 正 + 0.06 覧 + 0.06 出 + 0.06 降 + 0.06 残 + 0.05 接 - 0.05 需 - 0.05 規 + 0.05 索 - 0.05 包 + 0.05 険 + 0.05 奇 + 0.05 強 - 0.05 訴 - 0.05 戸 - 0.05 避 - 0.05 浄 + 0.05 文 - 0.05 猛 - 0.05 難 + 0.05 得 + 0.05 獲 - 0.05 美 - 0.05 督 - 0.05 監 - 0.05 失 + 0.05 語 + 0.05 久 + 0.05 閲 + 0.05 該 + 0.04 明 - 0.04 誌 - 0.04 稼 + 0.04 己 + 0.04 図 + 0.04 削 - 0.04 求 - 0.04 怠 - 0.04 及 - 0.04 小 + 0.04 思 + 0.04 超 + 0.04 腐 + 0.04 兼 - 0.04 撮 - 0.04 慶 - 0.04 塾 - 0.04 炊 + 0.03 魅 + 0.03 朝 + 0.03 南 + 0.03 先 + 0.03 興 - 0.03 暖 - 0.03 午 - 0.03 覚 + 0.03 側 - 0.03 講 + 0.03 裕 - 0.03 是 + 0.03 開 + 0.03 供 - 0.03 未 + 0.03 必 + 0.03 挿 - 0.03 雑 - 0.03 今 + 0.03 改 + 0.03 形 + 0.03 幸 - 0.03 易 - 0.03 燥 - 0.03 持 - 0.03 飯 - 0.03 災 - 0.03 鍵 - 0.03 欧 - 0.03 非 - 0.02 里 + 0.02 公 + 0.02 呈 - 0.02 刺 - 0.02 戦 - 0.02 瓶 - 0.02 港 + 0.02 崎 - 0.02 顔 - 0.02 風 - 0.02 冷 + 0.02 報 - 0.02 基 - 0.02 瀬 + 0.02 走 - 0.02 渉 - 0.02 還 - 0.02 盟 - 0.02 億 - 0.02 念 - 0.02 伴 + 0.02 種 - 0.02 症 + 0.02 川 + 0.01 称 - 0.01 朗 + 0.01 取 + 0.01 替 - 0.01 猫 + 0.01 苦 + 0.01 示 + 0.01 雲 + 0.01 著 + 0.01 哲 - 0.01 焦

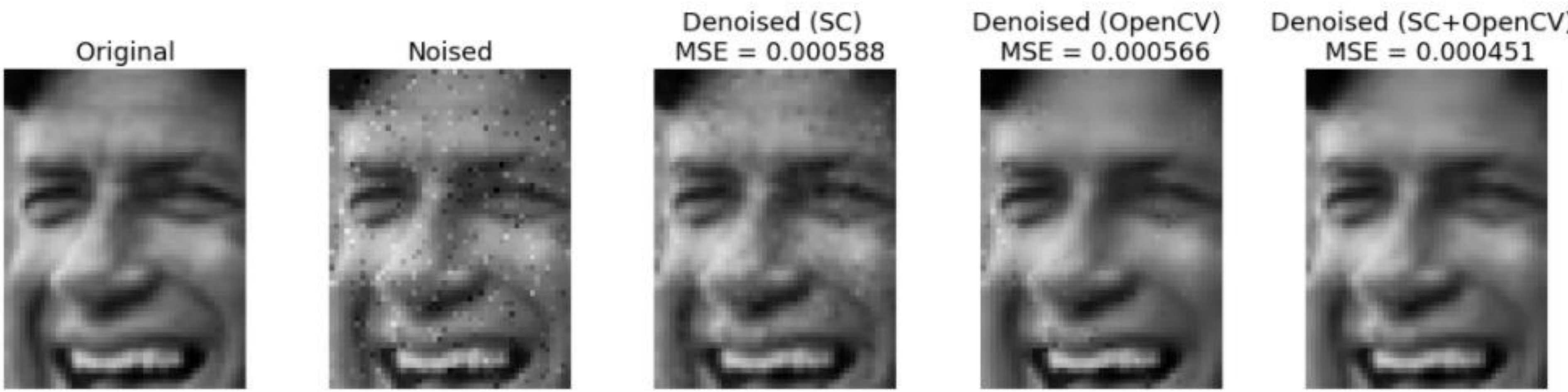
Sparse Codingによるノイズ除去

- 画像のノイズ除去
  - SC： L1正則化を用いたノイズ除去
  - CV： OpenCVを用いたノイズ除去
  - SC+CV： SCにさらにCVを適用

ノイズを粗く除去できる

画像がのっぺりしてしまう

ノイズを粗く除去してから、
少しだけのっぺりさせる



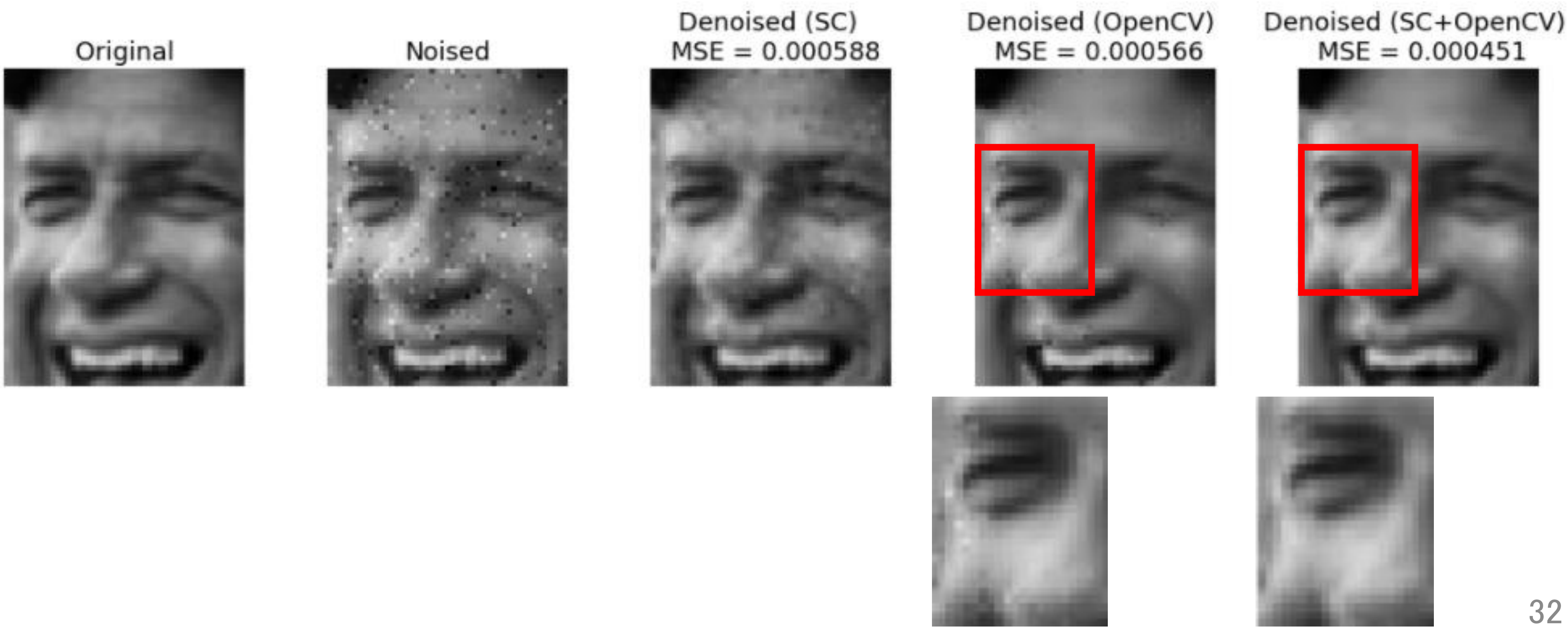
Sparse Codingによるノイズ除去

- 画像のノイズ除去
  - SC：L1正則化を用いたノイズ除去
  - CV：OpenCVを用いたノイズ除去
  - SC+CV：SCにさらにCVを適用

ノイズを粗く除去できる

画像がのっぺりしてしまう

ノイズを粗く除去してから、
少しだけのっぺりさせる



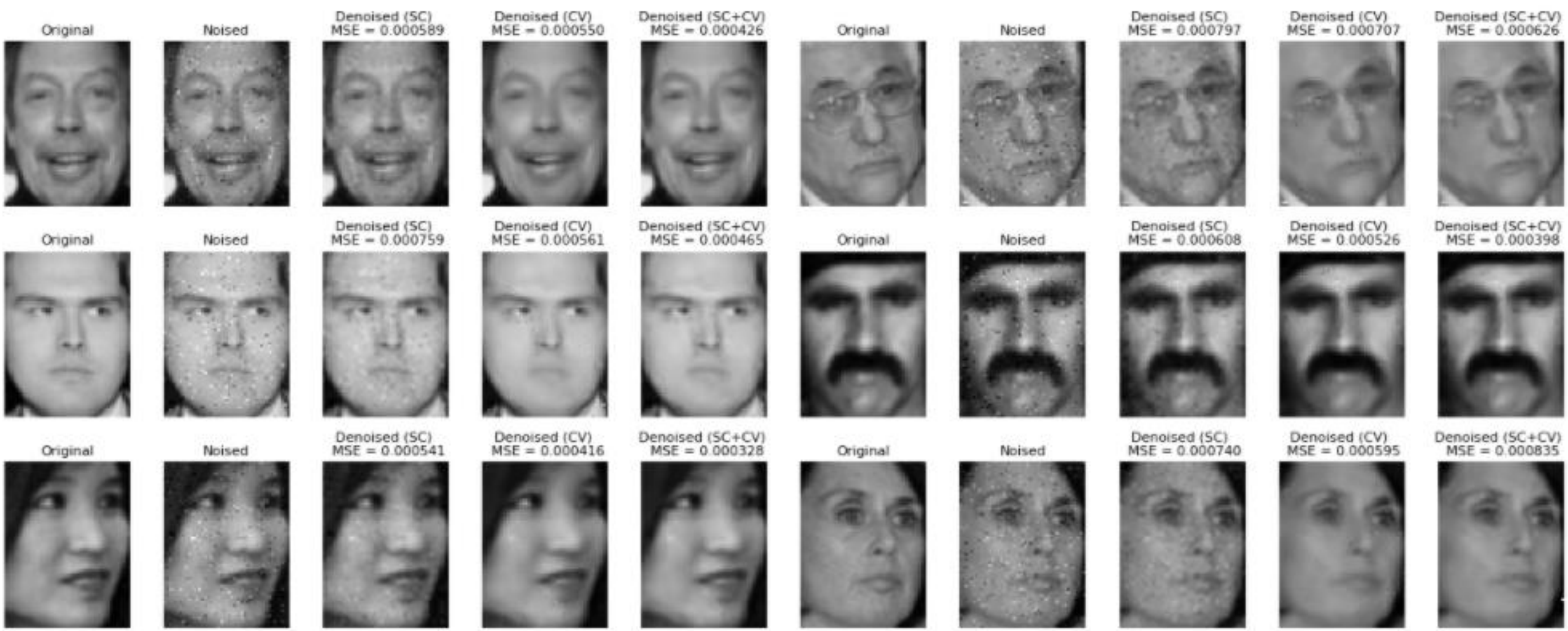
Sparse Codingによるノイズ除去

- 画像のノイズ除去
  - SC： L1正則化を用いたノイズ除去
  - CV： OpenCVを用いたノイズ除去
  - SC+CV： SCにさらにCVを適用

ノイズを粗く除去できる

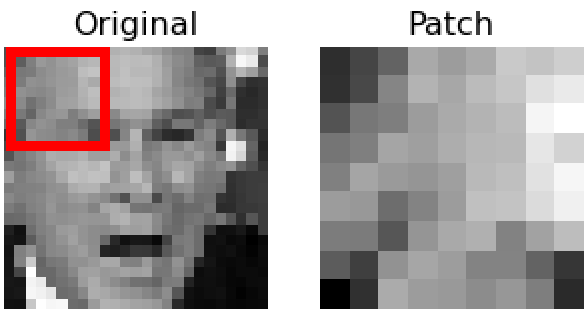
画像がのっぺりしてしまう

ノイズを粗く除去してから、
少しだけのっぺりさせる

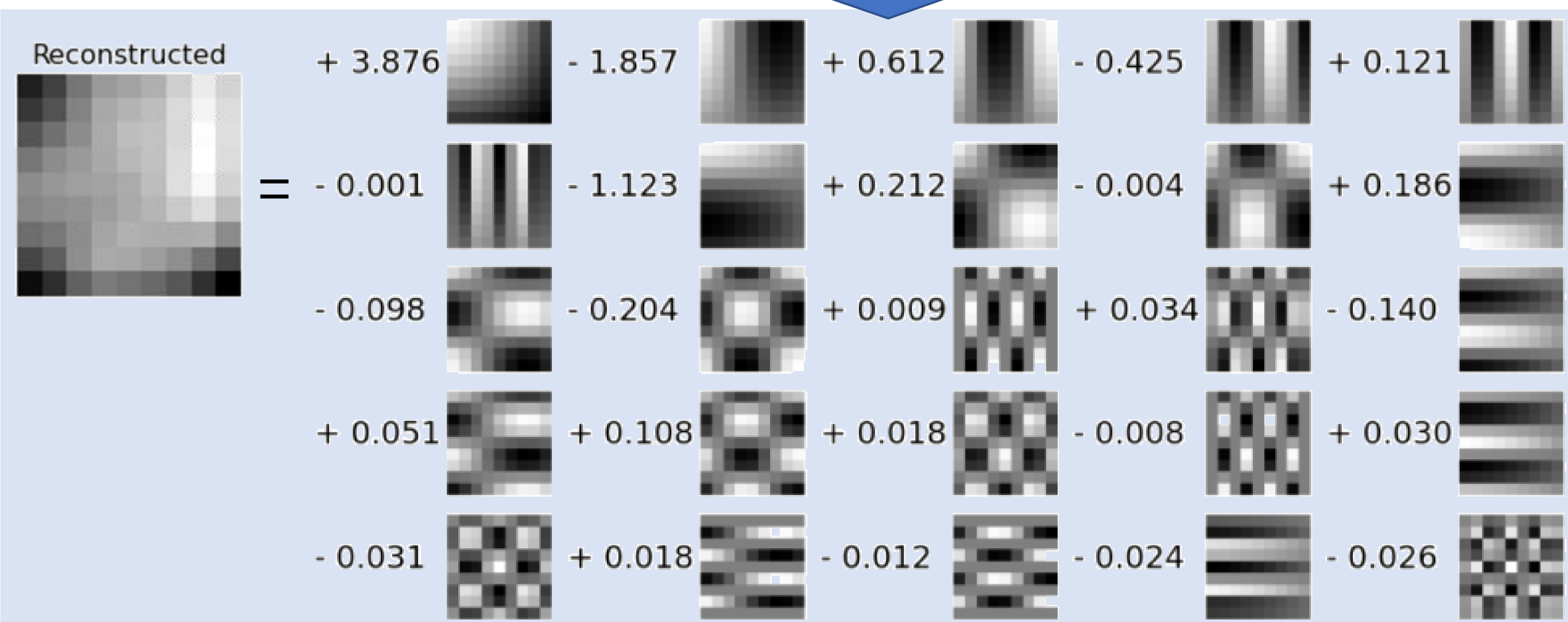


Sparse Codingによるノイズ除去のアイディア

- 画像を少数の基本パターンの和で表現する。

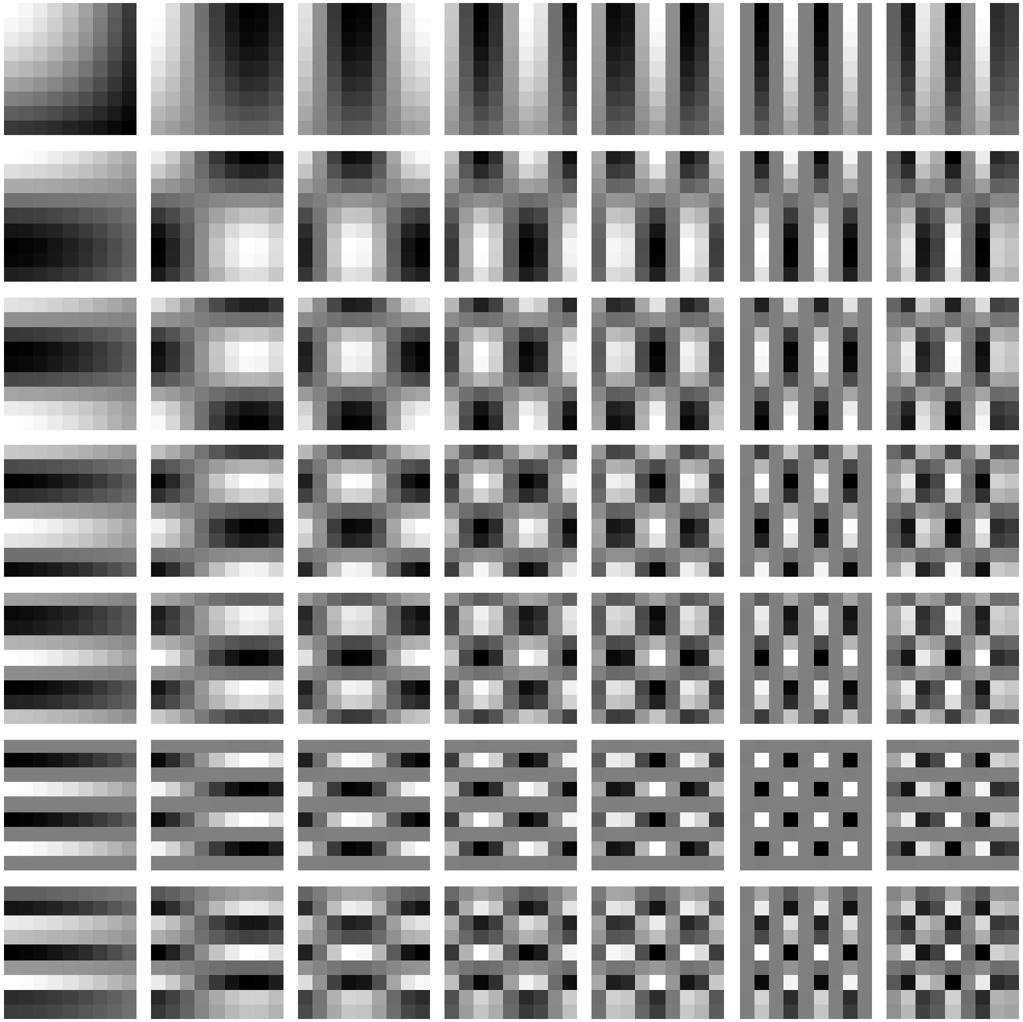


不規則なノイズを少数のパターンで表現するのは難しい。
↓
画像を少数のパターンの和で表現することで、ノイズ成分を除去出る。



【参考】離散コサイン変換の基底関数

- 画像の圧縮によく使われる基底関数。
  - 自然画像が高周波成分をほとんど含まない性質を利用。

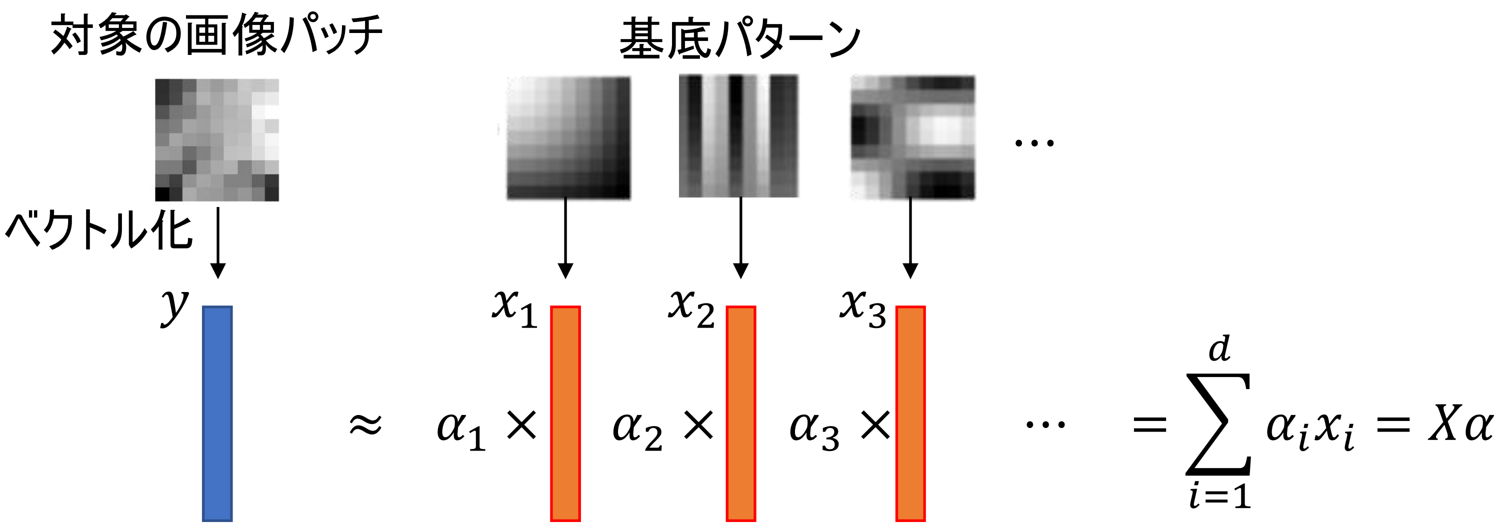


Sparse Codingによるノイズ除去のアイディア

- 複数の基底パターンから、画像の各パッチを表現するのに適切な少数のパターンを選ぶ。
  - 多数の基底パターンを使えばノイズも表現できてしまう。

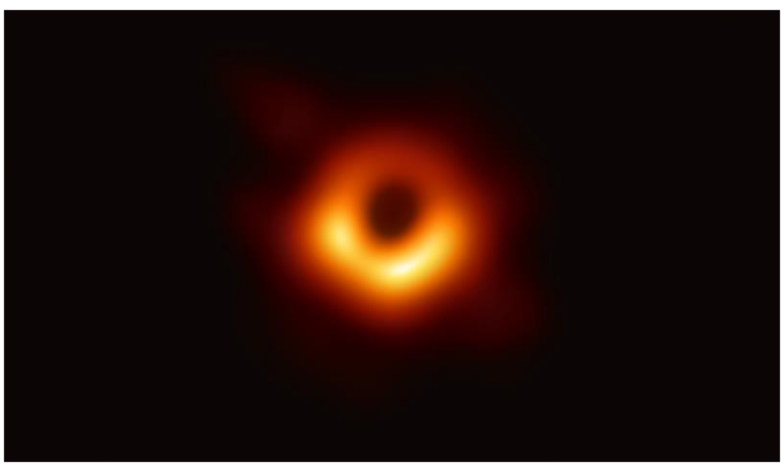
$$\min_{\alpha} \frac{1}{2}\|y - X\alpha\|^2 + \lambda\|\alpha\|_1$$

L1正則化のスパース性を使うことで少数の基底パターンだけを選び出す。



【参考】圧縮センシング

- 限られた観測から“背後の信号”の復元
- 標本化定理
  - “背後の信号”の復元には、その周波数の倍以上の周波数での観測が必要。
  - 一般に、観測が少ないと“背後の信号”は復元できない。
- 圧縮センシング
  - “背後の信号”がスパース（ゼロが多い）場合には、標本化定理よりも少ない観測から“背後の信号”が復元できる場合がある。



（圧縮センシングの例）
ブラックホールの撮像（2019/4/10）
https://eventhorizontelescope.org/press-release-april-10-2019-astronomers-capture-first-image-black-hole

教師あり学習の流れ

- データの用意
  - 入力xと出力yを決める。
  - 学習データ$D = \{x^{(n)}, y^{(n)}\}_{n=1}^{N}$を集める。
- モデルの学習
  - モデル候補（関数のクラス）を設定する。
  - 損失関数、正則化を設定する。
  - データにもっとも適合する関数をモデル候補の中から見つける。
  - 最適なハイパーパラメータを見つける。
- モデルの評価
  - 予測精度の評価

本日の講義内容

応用：特徴選択
Sparse Coding

L1正則化
グループ正則化

最適化
（近接勾配法）

グループ正則化

- 最適化したいパラメータにグループ構造がある場合
  - $y \approx x^\top \theta$
    - $x = \left[身長, 身長^2, \sqrt{身長}, 体重, 体重^2, \sqrt{体重}, 血圧, 血圧^2, \sqrt{血圧}\right]$
    - $\theta = [\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9]$
      - $\theta_1, \theta_2, \theta_3$は「身長」に関するパラメータ
      - $\theta_4, \theta_5, \theta_6$は「体重」に関するパラメータ
      - $\theta_7, \theta_8, \theta_9$は「血圧」に関するパラメータ

    - もしも「身長」が予測に関係ないなら、$\theta_1 = \theta_2 = \theta_3 = 0$となって欲しい。
      → グループ単位でパラメータがゼロになってくれると嬉しい。
      → グループ単位でのスパース性

    - $\theta_{G_1} = [\theta_1, \theta_2, \theta_3], \theta_{G_2} = [\theta_4, \theta_5, \theta_6], \theta_{G_3} = [\theta_7, \theta_8, \theta_9]$
      - $\theta_{G_1}, \theta_{G_2}, \theta_{G_3}$それぞれがゼロまたは非ゼロになって欲しい。

グループ正則化

- 最適化したいパラメータにグループ構造がある場合
  - $\theta_{G_1}, \theta_{G_2}, \ldots, \theta_{G_K}$をパラメータのグループとする。
    - 各グループそれぞれがゼロまたは非ゼロになって欲しい。

- グループ正則化
  - $\Omega(\theta) = \sum_{k=1}^{K} \|\theta_{G_k}\|$ L2ノルムを使う場合
  - $\Omega(\theta) = \sum_{k=1}^{K} \|\theta_{G_k}\|_\infty$ L∞ノルムを使う場合

$\|\theta_{G_k}\|^2$でないことに注意。二乗するとL2正則化になる。

- グループ正則化はL1正則化の一般化
  - グループのサイズが1（$\theta_{G_k} = \theta_k$）の場合
    - $\Omega(\theta) = \sum_{k=1}^{K} \|\theta_{G_k}\| = \sum_{k=1}^{K} |\theta_k|$

通常のL1正則化

本日の講義まとめ

- データの用意
  - 入力xと出力yを決める。
  - 学習データ$D = \left\{x^{(n)}, y^{(n)}\right\}_{n=1}^{N}$を集める。
- モデルの学習
  - モデル候補（関数のクラス）を設定する。
  - 損失関数、正則化を設定する。
  - データにもっとも適合する関数をモデル候補の中から見つける。
  - 最適なハイパーパラメータを見つける。
- モデルの評価
  - 予測精度の評価

本日の講義内容

応用：特徴選択
Sparse Coding

L1正則化
グループ正則化

最適化
（近接勾配法）

近接勾配法の収束性

- 収束性
  - 最適解： $\theta^* = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}}\, \phi(\theta) \coloneqq L(\theta) + \lambda \|\theta\|_1$
  - 更新式： $\theta^{[t+1]} = \mathrm{prox}_{\eta\lambda\|*\|_1}\left(\theta^{[t]} - \eta \nabla L\left(\theta^{[t]}\right)\right)$

$\nabla L\left(\theta^{[t]}\right) = \dfrac{\partial L\left(\theta^{[t]}\right)}{\partial \theta}$

  - 仮定
    - L は凸関数かつ微分可能
    - $\|\nabla L(\theta) - \nabla L(\theta')\| \leq M \|\theta - \theta'\|$ となる $0 < M < +\infty$ が存在する
    - $0 < \eta \leq \frac{1}{M}$

微分がLipschitz連続

  - このとき

$$\phi\left(\theta^{[T]}\right) - \phi(\theta^*) \leq \frac{1}{2\eta T} \left\|\theta^{[0]} - \theta^*\right\|^2 = O\left(\frac{1}{T}\right)$$

証明の流れ

- Lemma1 (Descent Lemma)

$$L(\theta') \leq L(\theta) + \nabla L(\theta)^\top(\theta' - \theta) + \frac{M}{2}\|\theta' - \theta\|^2, \ \forall\theta, \theta'$$

- Lemma 2
  - $G(\theta) = \frac{1}{\eta}\left(\theta - \mathrm{prox}_{\eta\lambda\|*\|_1}(\theta - \eta\nabla L(\theta))\right)$と定義するとき

$$\phi\left(\theta^{[t+1]}\right) \leq \phi(z) + G\left(\theta^{[t]}\right)^\top\left(\theta^{[t]} - z\right) - \frac{\eta}{2}\left\|G\left(\theta^{[t]}\right)\right\|^2, \ \forall z$$

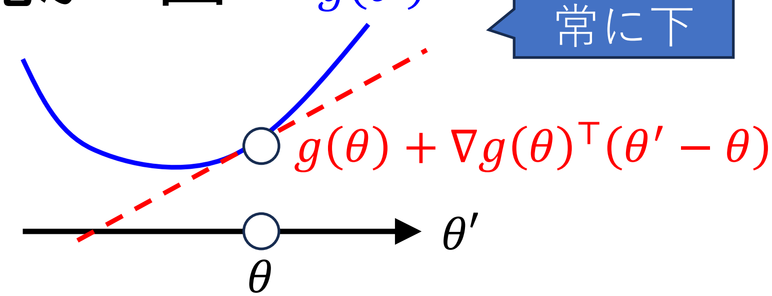
- 収束性

$$\phi\left(\theta^{[T]}\right) - \phi(\theta^*) \leq \frac{1}{2\eta T}\left\|\theta^{[0]} - \theta^*\right\|^2 = O\left(\frac{1}{T}\right)$$

準備

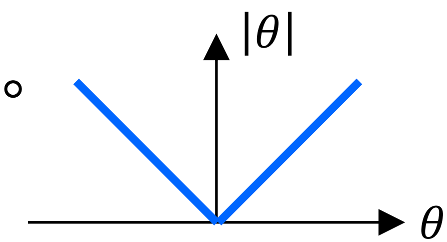
- 関数 $g(\theta)$ が θ について微分可能かつ凸

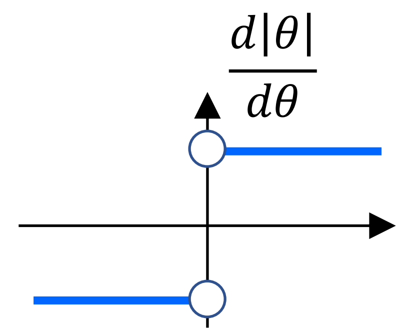
$$g(\theta') \geq g(\theta) + \nabla g(\theta)^\top(\theta' - \theta)$$



接線が常に下

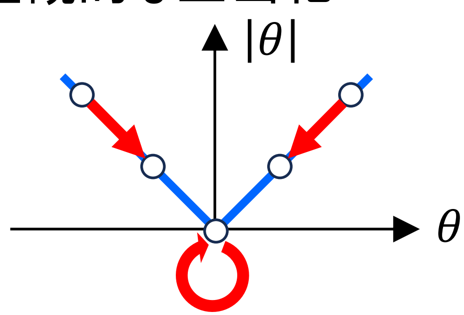
- 絶対値の“微分”
  - 絶対値は原点で微分できない。





  - 簡単のため **「原点での微分 = 0」** とする。
    - 直観的な正当化

厳密な正当化には“劣微分”を使う



$$\theta \leftarrow \theta - \eta \frac{d|\theta|}{d\theta}$$

勾配法による更新で
$|\theta|$の値が悪化しない

Lemma 1の証明

(1) L は微分可能, (2) $\exists M > 0, \|\nabla L(\theta) - \nabla L(\theta')\| \leq M\|\theta - \theta'\|$

このとき、$L(\theta') \leq L(\theta) + \nabla L(\theta)^\top(\theta' - \theta) + \frac{M}{2}\|\theta' - \theta\|^2, \ \ \forall\theta, \theta'$

$$L(\theta') - L(\theta) = \int_0^1 \nabla L\big(\theta + t(\theta' - \theta)\big)^\top (\theta' - \theta) dt$$

$$\frac{d}{dt} L\big(\theta + t(\theta' - \theta)\big) = \nabla L\big(\theta + t(\theta' - \theta)\big)^\top (\theta' - \theta)$$

$$= \nabla L(\theta)^\top(\theta' - \theta) + \int_0^1 \Big(\nabla L\big(\theta + t(\theta' - \theta)\big) - \nabla L(\theta)\Big)^\top (\theta' - \theta) dt$$

$$\leq \nabla L(\theta)^\top(\theta' - \theta) + \int_0^1 \Big\|\nabla L\big(\theta + t(\theta' - \theta)\big) - \nabla L(\theta)\Big\| \|\theta' - \theta\| dt$$

$$\leq \nabla L(\theta)^\top(\theta' - \theta) + \int_0^1 tM\|\theta' - \theta\|^2 dt$$

$$\leq \nabla L(\theta)^\top(\theta' - \theta) + \frac{M}{2}\|\theta' - \theta\|^2$$

（証明終了）

Lemma 2の証明

$G\left(\theta^{[t]}\right)=\frac{1}{\eta}\left(\theta^{[t]}-\theta^{[t+1]}\right)$ と定義するとき

$$\phi\left(\theta^{[t+1]}\right) \leq \phi(z)+G\left(\theta^{[t]}\right)^{\top}\left(\theta^{[t]}-z\right)-\frac{\eta}{2}\left\|G\left(\theta^{[t]}\right)\right\|^{2}, \forall z$$

- Lは凸： $L(z) \geq L\left(\theta^{[t]}\right)+\nabla L\left(\theta^{[t]}\right)^{\top}\left(z-\theta^{[t]}\right)$
- Lemma 1：

$$L\left(\theta^{[t+1]}\right) \leq L\left(\theta^{[t]}\right)+\nabla L\left(\theta^{[t]}\right)^{\top}\left(\theta^{[t+1]}-\theta^{[t]}\right)+\frac{M}{2}\left\|\theta^{[t+1]}-\theta^{[t]}\right\|^{2}$$

$$\Downarrow$$

$$\begin{aligned}
&L\left(\theta^{[t+1]}\right) \\
&\leq L(z)-\nabla L\left(\theta^{[t]}\right)^{\top}\left(z-\theta^{[t]}\right)+\nabla L\left(\theta^{[t]}\right)^{\top}\left(\theta^{[t+1]}-\theta^{[t]}\right)+\frac{M}{2}\left\|\eta G\left(\theta^{[t]}\right)\right\|^{2} \\
&= L(z)-\nabla L\left(\theta^{[t]}\right)^{\top}\left(z-\theta^{[t+1]}\right)+\frac{M}{2}\left\|\eta G\left(\theta^{[t]}\right)\right\|^{2}
\end{aligned}$$

①

Lemma 2の証明

$G(\theta^{[t]}) = \frac{1}{\eta}(\theta^{[t]} - \theta^{[t+1]})$ と定義するとき

$$\phi(\theta^{[t+1]}) \le \phi(z) + G(\theta^{[t]})^\top(\theta^{[t]} - z) - \frac{\eta}{2}\left\|G(\theta^{[t]})\right\|^2, \forall z$$

- $\|*\|_1$は凸： $\|z\|_1 \ge \left\|\theta^{[t+1]}\right\|_1 + \nabla\left\|\theta^{[t+1]}\right\|_1^\top(z - \theta^{[t+1]})$ ②
- $\theta^{[t+1]} = \underset{\theta}{\operatorname{argmin}} \frac{1}{2}\left\|\theta - \theta^{[t]} + \eta\nabla L(\theta^{[t]})\right\|^2 + \eta\lambda\|\theta\|_1$ の最適性条件

$$\theta^{[t+1]} - \theta^{[t]} + \eta\nabla L(\theta^{[t]}) + \eta\lambda\nabla\left\|\theta^{[t+1]}\right\|_1 = 0$$

$$\Rightarrow \lambda\nabla\left\|\theta^{[t+1]}\right\|_1 = -\frac{1}{\eta}\left(\theta^{[t+1]} - \theta^{[t]} + \eta\nabla L(\theta^{[t]})\right) = G(\theta^{[t]}) - \nabla L(\theta^{[t]})$$

③

Lemma 2の証明

$G\left(\theta^{[t]}\right) = \frac{1}{\eta}\left(\theta^{[t]} - \theta^{[t+1]}\right)$と定義するとき

$$\phi\left(\theta^{[t+1]}\right) \le \phi(z) + G\left(\theta^{[t]}\right)^{\top}\left(\theta^{[t]} - z\right) - \frac{\eta}{2}\left\|G\left(\theta^{[t]}\right)\right\|^2, \forall z$$

- ① & ②

$$L\left(\theta^{[t+1]}\right) + \lambda\left\|\theta^{[t+1]}\right\|_1$$
$$\le L(z) + \lambda\|z\|_1 - \nabla L\left(\theta^{[t]}\right)^{\top}\left(z - \theta^{[t+1]}\right) + \frac{M}{2}\left\|\eta G\left(\theta^{[t]}\right)\right\|^2 - \lambda\nabla\left\|\theta^{[t+1]}\right\|_1^{\top}\left(z - \theta^{[t+1]}\right)$$

- ③ & $\eta \le 1/M$

$$\phi\left(\theta^{[t+1]}\right)$$
$$\le \phi(z) - \nabla L\left(\theta^{[t]}\right)^{\top}\left(z - \theta^{[t+1]}\right) + \frac{M}{2}\left\|\eta G\left(\theta^{[t]}\right)\right\|^2 - \left(G\left(\theta^{[t]}\right) - \nabla L\left(\theta^{[t]}\right)\right)^{\top}\left(z - \theta^{[t+1]}\right)$$
$$= \phi(z) - G\left(\theta^{[t]}\right)^{\top}\left(z - \theta^{[t+1]}\right) + \frac{M}{2}\left\|\eta G\left(\theta^{[t]}\right)\right\|^2 = \phi(z) - G\left(\theta^{[t]}\right)^{\top}\left(z - \theta^{[t]} + \eta G\left(\theta^{[t]}\right)\right) + \frac{M}{2}\left\|\eta G\left(\theta^{[t]}\right)\right\|^2$$
$$= \phi(z) - G\left(\theta^{[t]}\right)^{\top}\left(z - \theta^{[t]}\right) - \eta\left\|G\left(\theta^{[t]}\right)\right\|^2 + \frac{M}{2}\left\|\eta G\left(\theta^{[t]}\right)\right\|^2 \le \phi(z) - G\left(\theta^{[t]}\right)^{\top}\left(z - \theta^{[t]}\right) - \frac{\eta}{2}\left\|G\left(\theta^{[t]}\right)\right\|^2$$

（証明終了）

収束性の証明

- Lemma 2において$z = \theta^{[t]}$とおく。
  - $\phi\left(\theta^{[t+1]}\right) \leq \phi\left(\theta^{[t]}\right) + G\left(\theta^{[t]}\right)^{\top}\left(\theta^{[t]} - \theta^{[t]}\right) - \frac{\eta}{2}\left\|G\left(\theta^{[t]}\right)\right\|^2$
    $= \phi\left(\theta^{[t]}\right) - \frac{\eta}{2}\left\|G\left(\theta^{[t]}\right)\right\|^2$

$$\Downarrow$$

  - $\phi\left(\theta^{[t+1]}\right) \leq \phi\left(\theta^{[t]}\right)$　　ϕの単調性

収束性の証明

- Lemma 2において$z = \theta^*$とおく。
  - $\phi\left(\theta^{[t+1]}\right) \le \phi(\theta^*) + G\left(\theta^{[t]}\right)^\top\left(\theta^* - \theta^{[t]}\right) - \frac{1}{2\eta}\left\|G\left(\theta^{[t]}\right)\right\|^2$

$$= \phi(\theta^*) + \frac{1}{2\eta}\left\|\theta^* - \theta^{[t]}\right\|^2 - \frac{1}{2\eta}\left\|\theta^* - \theta^{[t]} + \eta G\left(\theta^{[t]}\right)\right\|^2$$

$$= \phi(\theta^*) + \frac{1}{2\eta}\left\|\theta^* - \theta^{[t]}\right\|^2 - \frac{1}{2\eta}\left\|\theta^* - \theta^{[t+1]}\right\|^2$$

$$\sum_{t=0}^{T-1}\left(\phi\left(\theta^{[t+1]}\right) - \phi(\theta^*)\right) \ge \sum_{t=0}^{T-1}\left(\phi\left(\theta^{[T]}\right) - \phi(\theta^*)\right) = T\left(\phi\left(\theta^{[T]}\right) - \phi(\theta^*)\right)$$

ϕの単調性より

$$\sum_{t=0}^{T-1}\left(\phi\left(\theta^{[t+1]}\right) - \phi(\theta^*)\right) \le \frac{1}{2\eta}\sum_{t=0}^{T-1}\left(\left\|\theta^* - \theta^{[t]}\right\|^2 - \left\|\theta^* - \theta^{[t+1]}\right\|^2\right) \le \frac{1}{2\eta}\left\|\theta^* - \theta^{[0]}\right\|^2$$

（証明終了）

近接勾配法の収束性

- 収束性
  - 最適解: $\theta^* = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}}\, \phi(\theta) \coloneqq L(\theta) + \lambda \|\theta\|_1$
  - 更新式: $\theta^{[t+1]} = \operatorname{prox}_{\eta\lambda\|*\|_1}\left(\theta^{[t]} - \eta \nabla L\left(\theta^{[t]}\right)\right)$

$\nabla L\left(\theta^{[t]}\right) = \dfrac{\partial L\left(\theta^{[t]}\right)}{\partial \theta}$

  - 仮定
    - L は凸関数かつ微分可能
    - $\|\nabla L(\theta) - \nabla L(\theta')\| \leq M\|\theta - \theta'\|$ となる $0 < M < +\infty$ が存在する
    - $0 < \eta \leq \frac{1}{M}$

微分がLipschitz連続

  - このとき

$$\phi\left(\theta^{[T]}\right) - \phi(\theta^*) \leq \frac{1}{2\eta T}\left\|\theta^{[0]} - \theta^*\right\|^2 = O\left(\frac{1}{T}\right)$$

$\lambda = 0$とすれば通常の勾配法なので、勾配法についても$O\left(\frac{1}{T}\right)$が成り立つ。

そもそも普通の勾配法でも解けるのでは？

- **「原点での微分 = 0」**とするなら、普通の勾配でも解けるのでは？
  - 目的関数 $\phi(\theta) \coloneqq L(\theta) + \lambda\|\theta\|_1$
  - 勾配法 $\theta^{[t+1]} \leftarrow \theta^{[t]} - \eta\nabla\phi\left(\theta^{[t]}\right)$

厳密には“劣微分”を使った勾配法

- 実は普通の（劣）勾配法でも解ける。
- しかし、収束は一般に$O\left(\frac{1}{\sqrt{T}}\right)$であり、近接勾配法よりも遅い。

（劣）勾配法の収束性

- 収束性
  - 最適解： $\theta^* = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}}\, \phi(\theta)$
  - 更新式： $\theta^{[t+1]} = \theta^{[t]} - \eta \nabla \phi\left(\theta^{[t]}\right)$
  - 仮定
    - ϕ は凸関数
    - $\|\nabla \phi(\theta)\| \leq R$ となる $0 < R < +\infty$ が存在する　微分が有界
    - η を適切に選ぶ
  - このとき

$$\min_{t=0,1,\ldots,T} \phi\left(\theta^{[t]}\right) - \phi(\theta^*) = \frac{R}{\sqrt{T+1}} \left\|\theta^* - \theta^{[0]}\right\| = O\left(\frac{1}{\sqrt{T}}\right)$$

証明の流れ

- Lemma3

$$\phi\left(\theta^{[t]}\right) - \phi(\theta^*) \leq \frac{1}{2\eta}\left\|\theta^* - \theta^{[t]}\right\|^2 - \frac{1}{2\eta}\left\|\theta^* - \theta^{[t+1]}\right\|^2 + \frac{R^2}{2}\eta$$

$$\Downarrow$$

- 収束性

$$\min_{t=0,1,\ldots,T} \phi\left(\theta^{[t]}\right) - \phi(\theta^*) = O\left(\frac{1}{\sqrt{T}}\right)$$

Lemma 3の証明

$$\phi(\theta^{[t]}) - \phi(\theta^*) \le \frac{1}{2\eta}\left\|\theta^* - \theta^{[t]}\right\|^2 - \frac{1}{2\eta}\left\|\theta^* - \theta^{[t+1]}\right\|^2 + \frac{R^2}{2}\eta$$

- ϕ は凸 : $\phi(\theta^*) \ge \phi\left(\theta^{[t]}\right) + \nabla\phi\left(\theta^{[t]}\right)^\top\left(\theta^* - \theta^{[t]}\right)$

$$\phi\left(\theta^{[t]}\right) - \phi(\theta^*) \le \nabla\phi\left(\theta^{[t]}\right)^\top\left(\theta^{[t]} - \theta^*\right)$$

$$= \frac{1}{2\eta}\left\|\theta^* - \theta^{[t]}\right\|^2 - \frac{1}{2\eta}\left\|\theta^* - \theta^{[t+1]}\right\|^2 + \frac{\eta}{2}\left\|\nabla\phi\left(\theta^{[t]}\right)\right\|^2 \quad (*)$$

$$\le \frac{1}{2\eta}\left\|\theta^* - \theta^{[t]}\right\|^2 - \frac{1}{2\eta}\left\|\theta^* - \theta^{[t+1]}\right\|^2 + \frac{R^2}{2}\eta \quad \|\nabla\phi(\theta)\| \le R$$

$(*)$
$$\frac{1}{2\eta}\left\|\theta^* - \theta^{[t]}\right\|^2 - \frac{1}{2\eta}\left\|\theta^* - \theta^{[t+1]}\right\|^2 + \frac{\eta}{2}\left\|\nabla\phi\left(\theta^{[t]}\right)\right\|^2$$
$$= \frac{1}{2\eta}\left\|\theta - \theta^{[t]}\right\|^2 - \frac{1}{2\eta}\left\|\theta - \theta^{[t]} + \eta\nabla\phi\left(\theta^{[t]}\right)\right\|^2 + \frac{\eta}{2}\left\|\nabla\phi\left(\theta^{[t]}\right)\right\|^2$$
$$= -\nabla\phi\left(\theta^{[t]}\right)^\top\left(\theta - \theta^{[t]}\right)$$

（証明終了）

収束性の証明

$$\sum_{t=0}^{T}\left(\phi(\theta^{[t]})-\phi(\theta^*)\right)$$
$$\geq (T+1)\left(\min_{t=0,1,\ldots,T}\phi(\theta^{[t]})-\phi(\theta^*)\right)$$

$$\sum_{t=0}^{T}\left(\phi(\theta^{[t]})-\phi(\theta^*)\right)$$
$$\leq \sum_{t=0}^{T}\left(\frac{1}{2\eta}\left\|\theta^*-\theta^{[t]}\right\|^2-\frac{1}{2\eta}\left\|\theta^*-\theta^{[t+1]}\right\|^2+\frac{R^2}{2}\eta\right)$$
$$\leq \frac{1}{2\eta}\left\|\theta^*-\theta^{[0]}\right\|^2+\frac{(T+1)R^2}{2}\eta$$

$$\min_{t=0,1,\ldots,T}\phi(\theta^{[t]})-\phi(\theta^*)\leq\frac{1}{2(T+1)\eta}\left\|\theta^*-\theta^{[0]}\right\|^2+\frac{R^2}{2}\eta$$

- ηを適切に選んで右辺を小さくする。
  - $\eta=\frac{\|\theta^*-\theta^{[0]}\|}{R\sqrt{T+1}}$で右辺が最小。

$$\min_{t=0,1,\ldots,T}\phi(\theta^{[t]})-\phi(\theta^*)\leq\frac{R}{\sqrt{T+1}}\left\|\theta^*-\theta^{[0]}\right\|=O\left(\frac{1}{\sqrt{T}}\right)$$

（証明終了）