Stochastic Optimization
Introduction + Sparse regularization + Convex analysis

[†] [‡] Taiji Suzuki

[†]Tokyo Institute of Technology
Graduate School of Information Science and Engineering
Department of Mathematical and Computing Sciences
[‡]JST, PRESTO

Intensive course @ Nagoya University

Outline

1. Introduction
2. Short course to convex analysis
   - Convexity and related concepts
   - Duality
   - Smoothness and strong convexity

Lecture plan

- Day 1:
  - Convex analysis
  - First order method
  - "Online" stochastic optimization method: SGD, SRDA
- Day 2:
  - AdaGrad, acceleration of SGD
  - "Batch" stochastic optimization method: SDCA, SVRG, SAG
  - Distributed optimization (if possible)

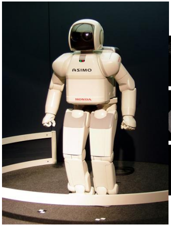
Outline

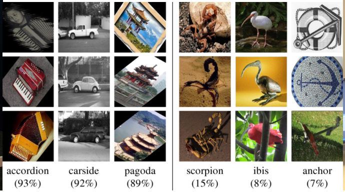
1. Introduction

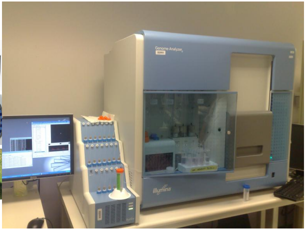
2. Short course to convex analysis
   - Convexity and related concepts
   - Duality
   - Smoothness and strong convexity

Machine learning as optimization

Machine learning is a methodology to deal with a lot of uncertain data.







Generalization error minimization
$$\min_{\theta \in \Theta} \mathrm{E}_Z[\ell_\theta(Z)]$$

Empirical approximation
$$\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell_\theta(z_i)$$

Stochastic optimization is an intersection of **learning** and **optimization**.

Recently stochastic optimization is used to treat **huge data**.

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell_\theta(z_i)}_{\textbf{Huge}} + \psi(\theta)$$

How to optimize this in efficient way?

Do we need to go through the whole data at every iteration?

History of stochastic optimization for ML

1951	**Robbins and Monro**	**Stochastic approximation** for root finding problem
1957	Rosenblatt	Perceptron
1978 1983	Nemirovskii and Yudin	Robustification for non-smooth obj. and optimality
1988 1992	Ruppert Polyak and Juditsky	Robust step size policy and averaging for smooth obj.
1998 2004	Bottou Bottou and LeCun	Online stochastic optimization **for large scale ML task**
2009-2012	Singer and Duchi; Duchi et al.; Xiao	FOBOS, AdaGrad, RDA
2012-2013	Le Roux et al. Shalev-Shwartz and Zhang Johnson and Zhang	Linear convergence on batch data (SAG,SDCA,SVRG)

Overview of stochastic optimization

$$\min_x f(x)$$

- Stochastic approximation (SA)
  - Optimization for systems with uncertainty,
    e.g., machine control, traffic management, social science, and so on.
  - $g_t = \nabla f(x^{(t)}) + \xi_t$ is observed where ξ_t is noise (typically i.i.d.).
- Stochastic approximation for machine learning and statistics
  - Typically generalization error minimization:

    $$\min_x f(x) = \min_x \mathrm{E}_Z[\ell(Z, x)].$$

  - $\ell(z, x)$ is a loss function:
    e.g., logistic loss $\ell((w, y), x) = \log(1 + \exp(-yw^\top x))$ for
    $z = (w, y) \in \mathbb{R}^p \times \{\pm 1\}$.
  - $g_t = \nabla \ell(z_t, x^{(t)})$ is observed where $z_t \sim P(Z)$ is i.i.d. data.
  - Used for **huge dataset**.
  - **We don't need exact optimization**. Optimization with certain precision (typically $O(1/n)$) is sufficient.

Two types of stochastic optimization

- Online type stochastic optimization:
  - We observe data sequentially.
  - Each observation is used just once (basically).

$$\min_x \ \mathrm{E}_Z[\ell(Z, x)]$$

- Batch type stochastic optimization
  - The whole sample has been already observed.
  - We may use training data multiple times.

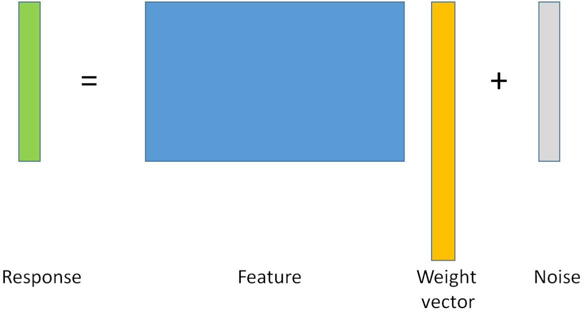
$$\min_x \ \frac{1}{n}\sum_{i=1}^{n}\ell(z_i, x)$$

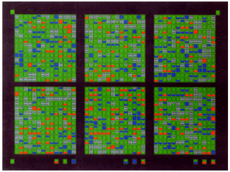
Summary of convergence rates

- Online methods (expected risk minimization):
  - $\frac{GR}{\sqrt{T}}$ (non-smooth, non-strongly convex)
  - $\frac{G^2}{\mu T}$ (non-smooth, strongly convex)
  - $\frac{\sigma R}{\sqrt{T}} + \frac{R^2 L}{T^2}$ (smooth, non-strongly convex)
  - $\frac{\sigma^2}{\mu T} + \exp\left(-\sqrt{\frac{\mu}{L}}T\right)$ (smooth, strongly convex)
- Batch methods (empirical risk minimization)
  - $\exp\left(-\frac{1}{n+\frac{\mu}{L}}T\right)$ (smooth loss, strongly convex reg)
  - $\exp\left(-\frac{1}{n+\sqrt{\frac{n\mu}{L}}}T\right)$ (smooth loss, strongly convex reg with acceleration)

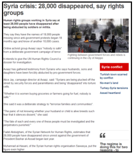
G: upper bound of norm of gradient, R: diameter of the domain,
L: smoothness, μ: strong convexity, σ: variance of the gradient

Example of empirical risk minimization: High dimensional data analysis





Bio-informatics


Syria crisis: 28,000 disappeared, say rights groups

Text data

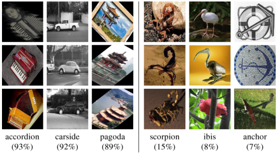
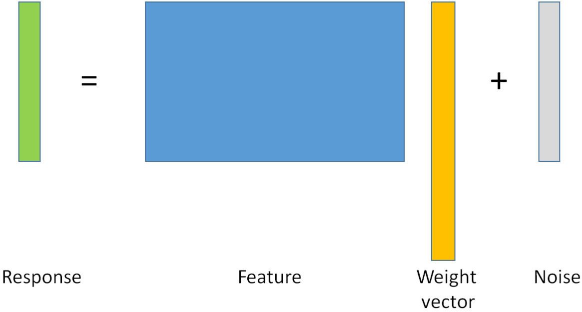
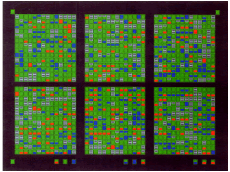


Image data

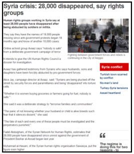
Example of empirical risk minimization: High dimensional data analysis



Redundant information deteriorates the estimation accuracy.



Bio-informatics



Syria crisis: 28,000 disappeared, say rights groups

Human rights groups working in Syria say at least 28,000 people have disappeared after being abducted by soldiers or militia.

They say they have the names of 18,000 people missing since anti-government protests began 18 months ago and know of another 10,000 cases.

Online activist group Avaaz says "nobody is safe" from a deliberate government campaign of terror.

It intends to give the UN Human Rights Council a dossier for investigation.

Fighting between government forces and rebels is continuing in the city of Aleppo

Avaaz has gathered testimony from Syrians who says husbands, sons and daughters have been forcibly abducted by pro-government forces.

Alice Jay, campaign director at Avaaz, said: "Syrians are being plucked off the street by security forces and paramilitaries and being 'disappeared' into torture cells.

"Whether it is women buying groceries or farmers going for fuel, nobody is safe."

She said it was a deliberate strategy to "terrorise families and communities".

"The panic of not knowing whether your husband or child is alive breeds such fear that it silences dissent," she said.

"The fate of each and every one of these people must be investigated and the perpetrators punished."

Fadel Abdulghani, of the Syrian Network for Human Rights, estimates that 28,000 people have disappeared since unrest against the government of President Bashar al-Assad began last year.

Muhannad al-Hasani, of the Syrian human rights organisation Sawasya, put the figure even higher.

Syria conflict

Turkish town scarred by conflict

No-man's land

Turkey-Syria tensions

Assad heartland

The regime is doing this for two reasons - to

Text data

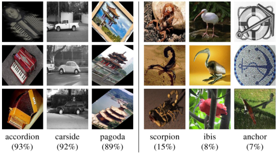
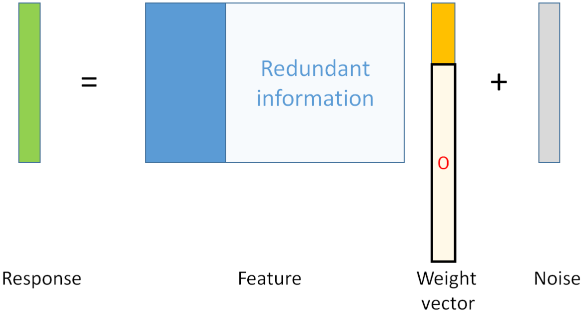


Image data

Sparse estimation



Cut off redundant information → sparsity

R. Tsibshirani (1996). Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B., Vol. 58, No. 1, pages 267–288.

Variable selection (linear regression)

Design matrix $X = (X_{ij}) \in \mathbb{R}^{n \times p}$.
p (**dimension**) $\gg n$ (**number of samples**).
The true vector $\beta^* \in \mathbb{R}^p$: <u>At most d non-zero elements (sparse)</u>.

$$\text{Linear model :} \quad Y = X\beta^* + \boldsymbol{\xi}.$$

Estimate β^* from (Y, X).
The number of parameters that we need to estimate is $d \rightarrow$ **variable selection**.

Variable selection (linear regression)

Design matrix $X = (X_{ij}) \in \mathbb{R}^{n \times p}$.
p (**dimension**) $\gg n$ (**number of samples**).
The true vector $\beta^* \in \mathbb{R}^p$: At most d non-zero elements (sparse).

$$\text{Linear model}: \quad Y = X\beta^* + \boldsymbol{\xi}.$$

Estimate β^* from (Y, X).
The number of parameters that we need to estimate is $d \to$ **variable selection**.

AIC:

$$\hat{\beta}_{\mathrm{AIC}} = \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \|Y - X\beta\|^2 + 2\sigma^2 \|\beta\|_0$$

where $\|\beta\|_0 = |\{j \mid \beta_j \neq 0\}|$.
$\to 2^p$ candidates. **NP-hard** $\quad \to$ **Convex approximation.**

Lasso estimator

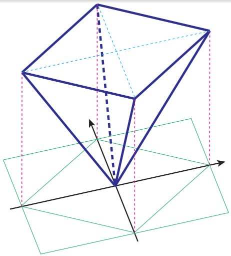
Lasso [L_1 regularization]

$$\hat{\beta}_{\text{Lasso}} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|^2 + \lambda \|\beta\|_1$$

where $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$.

→ **Convex optimization** !

- L_1-norm is the **convex hull** of L_0-norm on $[-1, 1]^p$ (the largest convex function which supports from below).
- L_1-norm is the **Lovász extension** of the cardinality function.



Lasso estimator

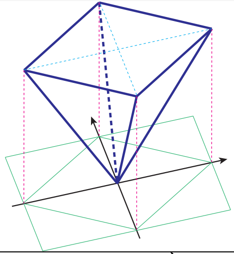
Lasso [L_1 regularization]

$$\hat{\beta}_{\mathrm{Lasso}} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|^2 + \lambda \|\beta\|_1$$

where $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$.

→ **Convex optimization** !

- L_1-norm is the **convex hull** of L_0-norm on $[-1,1]^p$ (the largest convex function which supports from below).
- L_1-norm is the **Lovász extension** of the cardinality function.



More generally for a loss function ℓ (logistic loss, hinge loss, ...)

$$\min_x \left\{ \sum_{i=1}^{n} \ell(z_i, x) + \lambda \|x\|_1 \right\}$$

Sparsity of Lasso estimator

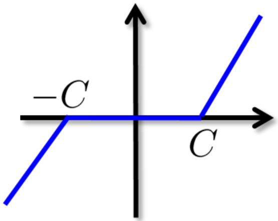
Suppose $p = n$ and $X = I$.

$$\hat{\beta}_{\mathrm{Lasso}} = \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \ \frac{1}{2}\|Y - \beta\|^2 + C\|\beta\|_1$$

$$\Rightarrow \quad \hat{\beta}_{\mathrm{Lasso},i} = \operatorname*{argmin}_{b \in \mathbb{R}} \ \frac{1}{2}(y_i - b)^2 + C|b|$$

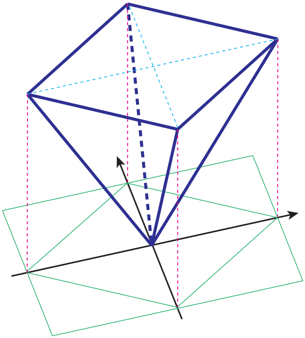
$$= \begin{cases} \mathrm{sign}(y_i)(y_i - C) & (|y_i| > C) \\ 0 & (|y_i| \le C). \end{cases}$$

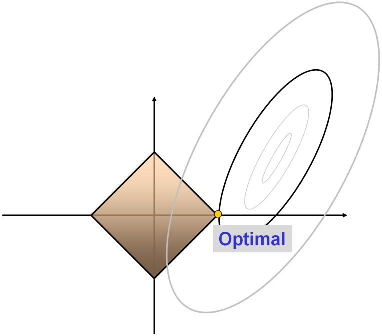
Small signal is shrunk to 0 → sparse !



Sparsity of Lasso estimator (fig)

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{n} \|X\beta - Y\|_2^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|.$$

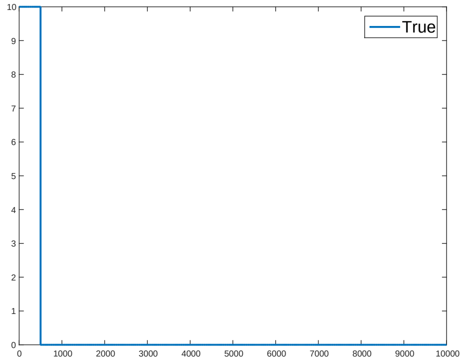




Example

$$Y = X\beta + \epsilon.$$

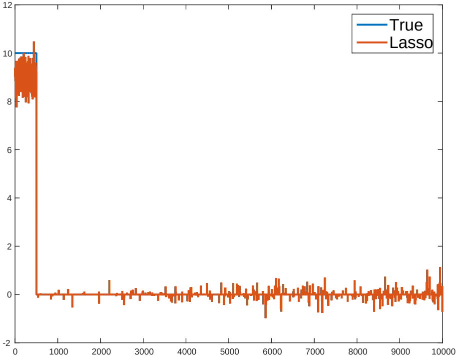
$$n = 1,000,\ p = 10,000,\ d = 500.$$



Example

$$Y = X\beta + \epsilon.$$

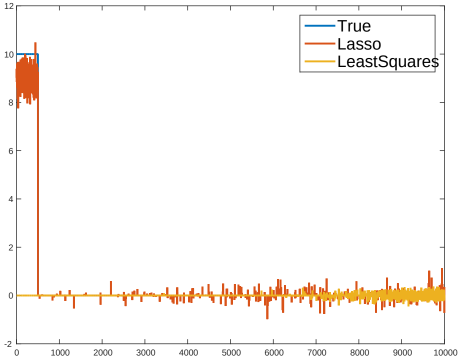
$$n = 1,000,\ p = 10,000,\ d = 500.$$



Example

$$Y = X\beta + \epsilon.$$

$$n = 1,000, \ p = 10,000, \ d = 500.$$



Benefit of sparsity

$$\hat{\beta} = \mathop{\arg\min}_{\beta \in \mathbb{R}^p} \frac{1}{n} \|X\beta - Y\|_2^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|.$$

Theorem (Lasso's convergence rate)

Under some conditions, there exists a constant C such that

$$\|\hat{\beta} - \beta^*\|_2^2 \leq C \frac{d \log(p)}{n}.$$

※ The overall dimension p affects just in $O(\log(p))$!
The actual dimension d is dominant.

Extensions of sparse regularization

$$\hat{\beta} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \frac{1}{n} \|X\beta - Y\|_2^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|$$

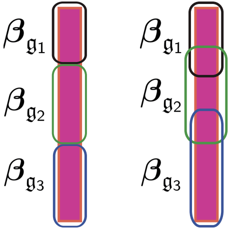
$$\downarrow$$

$$\hat{\beta} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \frac{1}{n} \ell(y_i, x_i^\top \beta) + \psi(\beta)$$

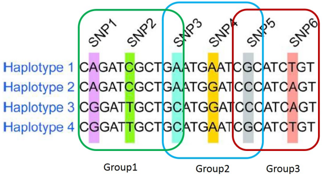
Examples

- Overlapped group lasso

$$\psi(\beta) = C \sum_{\mathfrak{g} \in \mathfrak{G}} \|\beta_{\mathfrak{g}}\|$$



- The groups may overlap.
- More aggressive sparsity.



Genome Wide Association Study (GWAS)

(Balding '06, McCarthy et al. '08)

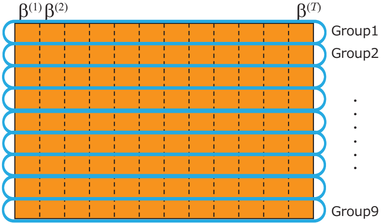
Application of group reg. (1)

- Multi-task learning (Lounici et al., 2009)

Estimate simultaneously across T tasks:

$$y_i^{(t)} = x_i^{(t)\top}\beta^{(t)} + \epsilon_i^{(t)} \quad (i = 1, \dots, n^{(t)},\ t = 1, \dots, T).$$

$$\min_{\beta^{(t)}} \quad \sum_{t=1}^{T}\sum_{i=1}^{n^{(t)}}(y_i - x_i^{(t)\top}\beta^{(t)})^2 + C\underbrace{\sum_{k=1}^{p}\|(\beta_k^{(1)}, \dots, \beta_k^{(T)})\|}_{\text{Group regularization}}.$$



Select non-zero elements across tasks

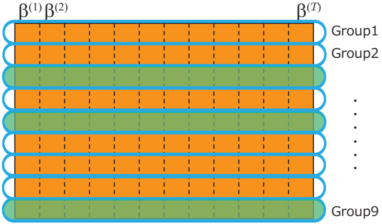
Application of group reg. (1)

- Multi-task learning (Lounici et al., 2009)

Estimate simultaneously across T tasks:

$$y_i^{(t)} = x_i^{(t)\top}\beta^{(t)} + \epsilon_i^{(t)} \quad (i = 1, \dots, n^{(t)},\ t = 1, \dots, T).$$

$$\min_{\beta^{(t)}} \quad \sum_{t=1}^{T}\sum_{i=1}^{n^{(t)}}(y_i - x_i^{(t)\top}\beta^{(t)})^2 + C\underbrace{\sum_{k=1}^{p}\|(\beta_k^{(1)}, \dots, \beta_k^{(T)})\|}_{\text{Group regularization}}.$$



Select non-zero elements across tasks

Application of group reg. (2)

- Sentence regularization for text classification (Yogatama and Smith, 2014)

The words occurred in the same sentence is grouped:

$$\psi(\beta) = \sum_{d=1}^{D} \sum_{s=1}^{S_d} \lambda_{d,s} \|\beta_{(d,s)}\|_2,$$

(d expresses a document, s expresses a sentence).

Table 4. A review from Amazon dvd review dataset categorized as a positive review. Each line is a sentence identified by the sentence segmenter. There are five sentences in this article. Selected sentences in the learner's copy variables are highlighted in **blue and bold**. We also display the color-coded log-odds scores, as discussed in the text (sentence, elastic, ridge, lasso) based on removing each sentence for each competing model. We only display scores that are greater than 10^{-3} in absolute values.

Sentence	Negative	Positive
this film is one big joke : you have all the basics elements		(0.42)
of romance (love at first sight , great passion , etc .) and gangster flicks		(0.22)
(brutality , dagerous machinations , the mysterious don , etc.) ,		(0.07)
but it is all done with the crudest humor .		(0.48)
it ' s the kind of thing you either like viserally and		(0.01)
immediately " get " or you don ' t .		(0.01)
that is a matter of taste and expectations .		(0.01)
i enjoyed it and it took me back to the mid80s ,		(0.02)
when nicolson and turner were in their primes .		(0.01)
the acting is very good , if a bit obviously tongue - in - cheek .		(0.01)

Trace norm regularization

W : $M \times N$ matrix.

$$\|W\|_{\mathrm{Tr}} = \mathrm{Tr}[(WW^\top)^{\frac{1}{2}}] = \sum_{j=1}^{\min\{M,N\}} \sigma_j(W)$$

$\sigma_j(W)$ is the j-th singular value of W (non-negative).

- Sum of singular values = L_1-regularization on singular values
  $\rightarrow$ Singular values are **sparse**
- Sparse singular values = **Low rank**

Application of trace norm reg.: Recommendation system

	Movie A	Movie B	Movie C	$\cdots$	Movie X
User 1	4	8	*	$\cdots$	2
User 2	2	*	2	$\cdots$	*
User 3	2	4	*	$\cdots$	*
$\vdots$					

(e.g., Srebro et al. (2005), NetFlix Bennett and Lanning (2007))

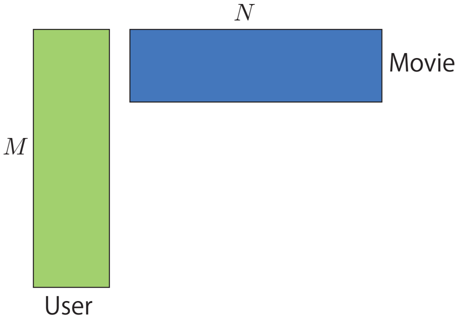
Application of trace norm reg.: Recommendation system

Assuming the rank is 1.

	Movie A	Movie B	Movie C	$\cdots$	Movie X
User 1	4	8	4	$\cdots$	2
User 2	2	4	2	$\cdots$	1
User 3	2	4	2	$\cdots$	1
$\vdots$					

(e.g., Srebro et al. (2005), NetFlix Bennett and Lanning (2007))

Application of trace norm reg.: Recommendation system



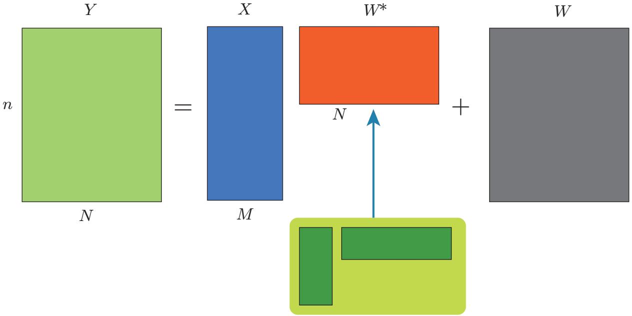
→ Low rank matrix completion:

- Rademacher complexity of low rank matrices: Srebro et al. (2005).
- Compressed sensing: Candès and Tao (2009), Candès and Recht (2009).

Example: Reduced rank regression

- Reduced rank regression (Anderson, 1951, Burket, 1964, Izenman, 1975)
- Multi-task learning (Argyriou et al., 2008)

Reduced rank regression

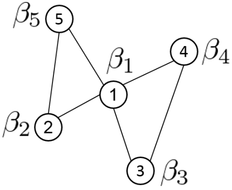


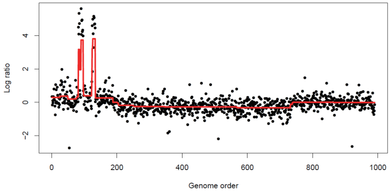
W^* is low rank.

(Generalized) Fused Lasso

$$\psi(\beta) = C \sum_{(i,j)\in E} |\beta_i - \beta_j|.$$

(Tibshirani et al. (2005), Jacob et al. (2009))





Genome data analysis by Fused lasso (Tibshirani and Taylor '11)

TV-denoising (Chambolle '04)

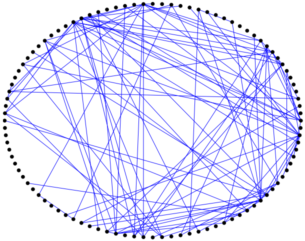
Sparse covariance selection

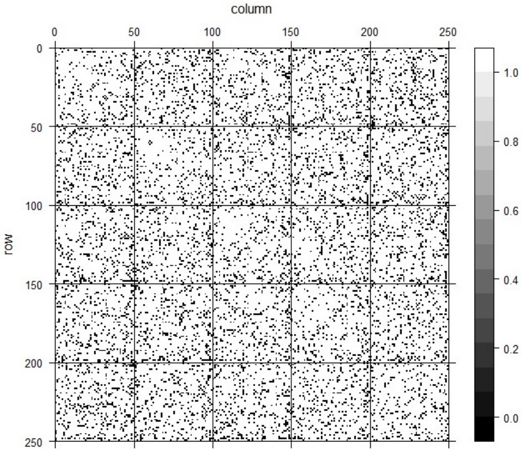
$x_k \sim N(0, \Sigma)$ (i.i.d., $\Sigma \in \mathbb{R}^{p \times p}$), $\quad \widehat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n} x_k x_k^\top$.

$$\hat{S} = \mathop{\mathrm{argmin}}_{S \succeq O} \left\{ -\log(\det(S)) + \mathrm{Tr}[S\widehat{\Sigma}] + \lambda \sum_{i,j=1}^{p} |S_{i,j}| \right\}.$$

(Meinshausen and B uhlmann, 2006, Yuan and Lin, 2007, Banerjee et al., 2008)

- Estimating the inverse S of Σ.
- $S_{i,j} = 0 \;\Leftrightarrow X_{(i)}, X_{(j)}$ is conditionally independent.
- Gaussian graphical model can be estimated by convex optimization.





Covariance selection on the stock data of 50 randomly selected companies in NASDAQ list from 4 January 2011 to 31 December 2014.
(Lie Michael, Bachelor thesis)

Other examples

- Robust PCA (Candés et al. 2009).
- Low rank tensor estimation (Signoretto et al., 2010; Tomioka et al., 2011).
- Dictionary learning (Kasiviswanathan et al., 2012; Rakotomamonjy, 2013).

Outline

1. Introduction
2. Short course to convex analysis
   - Convexity and related concepts
   - Duality
   - Smoothness and strong convexity

Regularized empirical risk minimization

Basically, we want to solve

- Empirical risk minimization:

$$\min_{x \in \mathbb{R}^p} \quad \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, x).$$

- Regularized empirical risk minimization:

$$\min_{x \in \mathbb{R}^p} \quad \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, x) + \psi(x).$$

In this lecture, we assume ℓ and ψ are **convex**.
$\rightarrow$ convex analysis to exploit the properties of convex functions.

Outline

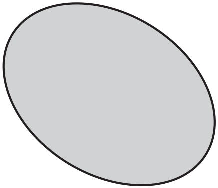
1. Introduction
2. Short course to convex analysis
   - Convexity and related concepts
   - Duality
   - Smoothness and strong convexity

Convex set

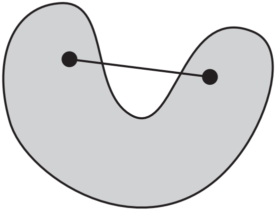
Definition (Convex set)

A convex set is a set that contains the segment connecting two points in the set:

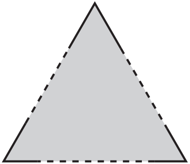
$$x_1, x_2 \in C \quad \Longrightarrow \quad \theta x_1 + (1 - \theta_2) x_2 \in C \quad (\theta \in [0,1]).$$



Convex set



Non-convex set



Non-convex set

Epigraph and domain

Let $\bar{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$.

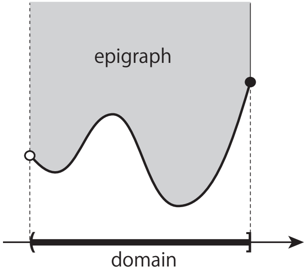
Definition (Epigraph and domain)

- The epigraph of a function $f : \mathbb{R}^p \to \bar{\mathbb{R}}$ is given by

$$\mathrm{epi}(f) := \{(x, \mu) \in \mathbb{R}^{p+1} : f(x) \le \mu\}.$$

- The domain of a function $f : \mathbb{R}^p \to \bar{\mathbb{R}}$ is given by

$$\mathrm{dom}(f) := \{x \in \mathbb{R}^p : f(x) < \infty\}.$$



Convex function

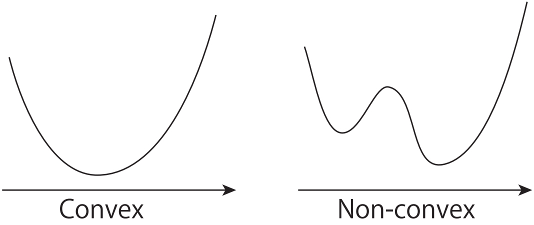
Let $\bar{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$.

Definition (Convex function)

A function $f : \mathbb{R}^p \to \bar{\mathbb{R}}$ is a convex function if f satisfies

$$\theta f(x) + (1-\theta) f(y) \geq f(\theta x + (1-\theta) y) \quad (\forall x, y \in \mathbb{R}^p, \theta \in [0,1]),$$

where $\infty + \infty = \infty$, $\infty \leq \infty$.



- f is convex $\Leftrightarrow$ $\mathrm{epi}(f)$ is a convex set.

Proper and closed convex function

- If the domain of a function f is not empty ($\mathrm{dom}(f) \neq \emptyset$), f is called proper.
- If the epigraph of a convex function f is a closed set, then f is called closed.

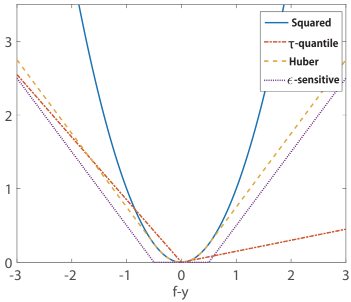
  (We are interested in only a proper closed function in this lecture.)

- Even if f is closed, it's domain is not necessarily closed (even for 1D).
- "f is closed" does not imply "f is continuous."
- Closed convex function is continuous on a <u>segment</u> in its domain.
- Closed function is "lower semicontinuity."

Convex loss functions (regression)

All well used loss functions are (closed) convex. The followings are convex w.r.t. u with a fixed label $y \in \mathbb{R}$.

- Squared loss: $\ell(y, u) = \frac{1}{2}(y - u)^2$.
- τ-quantile loss: $\ell(y, u) = (1 - \tau) \max\{u - y, 0\} + \tau \max\{y - u, 0\}$. for some $\tau \in (0, 1)$. Used for quantile regression.
- ϵ-sensitive loss: $\ell(y, u) = \max\{|y - u| - \epsilon, 0\}$ for some $\epsilon > 0$. Used for support vector regression.

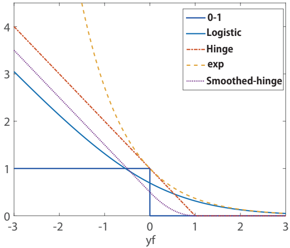


Convex surrogate loss (classification)

$y \in \{\pm 1\}$

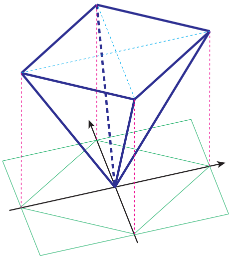
- Logistic loss: $\ell(y, u) = \log((1 + \exp(-yu))/2)$.
- Hinge loss: $\ell(y, u) = \max\{1 - yu, 0\}$.
- Exponential loss: $\ell(y, u) = \exp(-yu)$.
- Smoothed hinge loss:

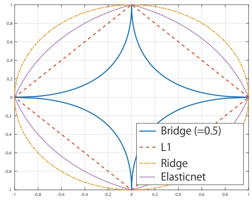
$$\ell(y, u) = \begin{cases} 0, & (yu \geq 1), \\ \frac{1}{2} - yu, & (yu < 0), \\ \frac{1}{2}(1 - yu)^2, & (\text{otherwise}). \end{cases}$$



Convex regularization functions

- Ridge regularization: $R(x) = \|x\|_2^2 := \sum_{j=1}^{p} x_j^2$.
- L_1 regularization: $R(x) = \|x\|_1 := \sum_{j=1}^{p} |x_j|$.
- Trace norm regularization: $R(X) = \|X\|_{\mathrm{tr}} = \sum_{k=1}^{\min\{q,r\}} \sigma_j(X)$ where $\sigma_j(X) \geq 0$ is the j-th singular value.





$\frac{1}{n}\sum_{i=1}^{n}(y_i - z_i^\top x)^2 + \lambda\|x\|_1$: Lasso
$\frac{1}{n}\sum_{i=1}^{n}\log(1+\exp(-y_i z_i^\top x)) + \lambda\|X\|_{\mathrm{tr}}$: Low rank matrix recovery

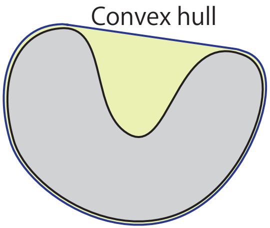
Other definitions of sets

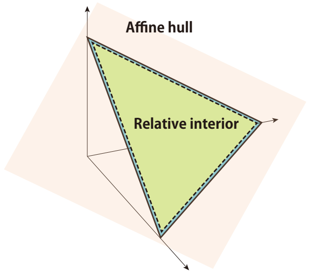
Convex hull: $\mathrm{conv}(C)$ is the smallest convex set that contains a set $C \subseteq \mathbb{R}^p$.

Affine set: A set A is an affine set if and only if $\forall x, y \in A$, the line that intersects x and y lies in A: $\lambda x + (1-\lambda)y \;\; \forall \lambda \in \mathbb{R}$.

Affine hull: The smallest affine set that contains a set $C \subseteq \mathbb{R}^p$.

Relative interior: $\mathrm{ri}(C)$. Let A be the affine hull of a convex set $C \subseteq \mathbb{R}^p$. $\mathrm{ri}(C)$ is a set of internal points with respect to the relative topology induced by the affine hull A.





Continuity of a closed convex function

Theorem

For a (possibly non-convex) function $f : \mathbb{R}^p \to \bar{\mathbb{R}}$, the following three conditions are equivalent to each other.

1. *f is lower semi-continuous.*
2. *For any converging sequence $\{x_n\}_{n=1}^{\infty} \subseteq \mathbb{R}^p$ s.t. $x_\infty = \lim_n x_n$, $\liminf_n f(x_n) \geq f(x_\infty)$.*
3. *f is closed.*

Remark: Any convex function f is continuous in $\mathrm{ri}(\mathrm{dom}(f))$. The continuity could be broken on the boundary of the domain.

Outline

1. Introduction

2. Short course to convex analysis
   - Convexity and related concepts
   - Duality
   - Smoothness and strong convexity

Subgradient

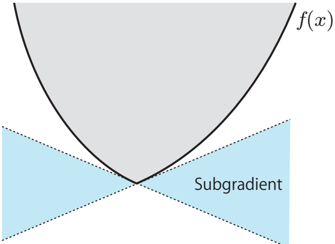
We want to deal with non-differentiable function such as L_1 regularization. To do so, we need to define something like gradient.

Definition (Subdifferential, subgradient)

For a proper convex function $f : \mathbb{R}^p \to \bar{\mathbb{R}}$, the **subdifferential** of f at $x \in \mathrm{dom}(f)$ is defined by

$$\partial f(x) := \{g \in \mathbb{R}^p \mid \langle x' - x, g\rangle + f(x) \le f(x') \ \ (\forall x' \in \mathbb{R}^p)\}.$$

An element of the subdifferential is called **subgradient**.



Properties of subgradient

- Subgradient does not necessarily exist ($\partial f(x)$ could be empty).
  $f(x) = x \log(x)$ $(x \geq 0)$ is proper convex but not subdifferentiable at $x = 0$.
- Subgradient always exists on $\mathrm{ri}(\mathrm{dom}(f))$.

Properties of subgradient

- Subgradient does not necessarily exist ($\partial f(x)$ could be empty).
  $f(x) = x\log(x)$ $(x \geq 0)$ is proper convex but not subdifferentiable at $x = 0$.
- Subgradient always exists on $\mathrm{ri}(\mathrm{dom}(f))$.
- If f is differentiable at x, its gradient is the unique element of subdiff.

$$\partial f(x) = \{\nabla f(x)\}.$$

Properties of subgradient

- Subgradient does not necessarily exist ($\partial f(x)$ could be empty).
  $f(x) = x\log(x)$ ($x \geq 0$) is proper convex but not subdifferentiable at $x = 0$.
- Subgradient always exists on $\mathrm{ri}(\mathrm{dom}(f))$.
- If f is differentiable at x, its gradient is the unique element of subdiff.

$$\partial f(x) = \{\nabla f(x)\}.$$

- If $\mathrm{ri}(\mathrm{dom}(f)) \cap \mathrm{ri}(\mathrm{dom}(h)) \neq \emptyset$, then

$$\begin{aligned}\partial(f+h)(x) &= \partial f(x) + \partial h(x) \\ &= \{g + g' \mid g \in \partial f(x),\ g' \in \partial h(x)\} \\ &\quad (\forall x \in \mathrm{dom}(f) \cap \mathrm{dom}(h)).\end{aligned}$$

Properties of subgradient

- Subgradient does not necessarily exist ($\partial f(x)$ could be empty).
  $f(x) = x\log(x)$ ($x \geq 0$) is proper convex but not subdifferentiable at $x = 0$.
- Subgradient always exists on $\mathrm{ri}(\mathrm{dom}(f))$.
- If f is differentiable at x, its gradient is the unique element of subdiff.
$$\partial f(x) = \{\nabla f(x)\}.$$
- If $\mathrm{ri}(\mathrm{dom}(f)) \cap \mathrm{ri}(\mathrm{dom}(h)) \neq \emptyset$, then
$$\begin{aligned}\partial(f+h)(x) &= \partial f(x) + \partial h(x) \\ &= \{g + g' \mid g \in \partial f(x),\ g' \in \partial h(x)\} \\ &\quad (\forall x \in \mathrm{dom}(f) \cap \mathrm{dom}(h)).\end{aligned}$$
- For all $g \in \partial f(x)$ and all $g' \in \partial f(x')$ ($x, x' \in \mathrm{dom}(f)$),
$$\langle g - g', x - x' \rangle \geq 0.$$

Legendre transform

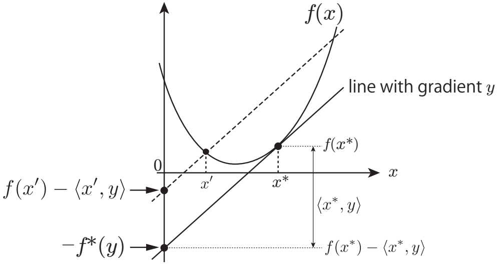
Defines the other representation on the **dual space** (the space of **gradients**).

Definition (Legendre transform)

Let f be a (possibly non-convex) function $f : \mathbb{R}^p \to \bar{\mathbb{R}}$ s.t. $\mathrm{dom}(f) \neq \emptyset$.
Its **convex conjugate** is given by

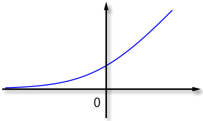
$$f^*(y) := \sup_{x \in \mathbb{R}^p} \{\langle x, y\rangle - f(x)\}.$$

The map from f to f^* is called Legendre transform.

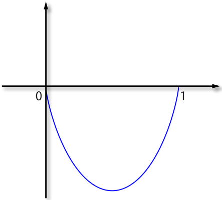


Examples

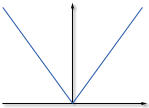
	$f(x)$	$f^*(y)$
Squared loss	$\frac{1}{2}x^2$	$\frac{1}{2}y^2$
Hinge loss	$\max\{1-x, 0\}$	$\begin{cases} y & (-1 \le y \le 0), \\ \infty & (\text{otherwise}). \end{cases}$
Logistic loss	$\log(1+\exp(-x))$	$\begin{cases} (-y)\log(-y) + (1+y)\log(1+y) & (-1 \le y \le 0), \\ \infty & (\text{otherwise}). \end{cases}$
L_1 regularization	$\\|x\\|_1$	$\begin{cases} 0 & (\max_j \|y_j\| \le 1), \\ \infty & (\text{otherwise}). \end{cases}$
L_p regularization ($p > 1$)	$\sum_{j=1}^d \|x_j\|^p$	$\sum_{j=1}^d \frac{p-1}{p^{\frac{p}{p-1}}} \|y_j\|^{\frac{p}{p-1}}$

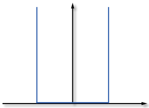


logistic



dual of logistic


L_1-norm


dual

Properties of Legendre transform

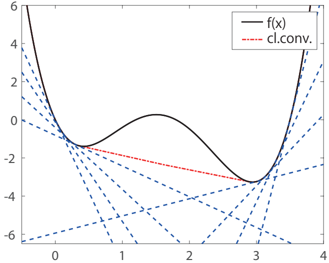
- f^* is a convex function even if f is not.
- f^{**} is the closure of the convex hull of f:

$$f^{**} = \mathrm{cl}(\mathrm{conv}(f)).$$

Corollary

Legendre transform is a *bijection* *from the set of proper closed convex functions onto that defined on the dual space.*

$$f \quad (\textit{proper closed convex}) \quad \Leftrightarrow \quad f^* \quad (\textit{proper closed convex})$$



Connection to subgradient

Lemma

$$y \in \partial f(x) \quad \Leftrightarrow \quad f(x) + f^*(y) = \langle x, y \rangle \quad \Leftrightarrow \quad x \in \partial f^*(y).$$

$$\begin{aligned} \because \quad y \in \partial f(x) &\Rightarrow x = \operatorname*{argmax}_{x' \in \mathbb{R}^p} \{\langle x', y \rangle - f(x')\} \\ &\qquad\qquad \text{(take the "derivative" of } \langle x', y \rangle - f(x')) \\ &\Rightarrow f^*(y) = \langle x, y \rangle - f(x). \end{aligned}$$

Remark: By definition, we always have

$$f(x) + f^*(y) \geq \langle x, y \rangle.$$

$\rightarrow$ **Young-Fenchel's inequality**.

★ Fenchel's duality theorem

Theorem (Fenchel's duality theorem)

Let $f : \mathbb{R}^p \to \bar{\mathbb{R}}$, $g : \mathbb{R}^q \to \bar{\mathbb{R}}$ be proper closed convex, and $A \in \mathbb{R}^{q \times p}$. Suppose that either of condition (a) or (b) is satisfied, then it holds that

$$\inf_{x \in \mathbb{R}^p} \{f(x) + g(Ax)\} = \sup_{y \in \mathbb{R}^q} \{-f^*(A^\top y) - g^*(-y)\}.$$

(a) $\exists x \in \mathbb{R}^p$ s.t. $x \in \mathrm{ri}(\mathrm{dom}(f))$ and $Ax \in \mathrm{ri}(\mathrm{dom}(g))$.

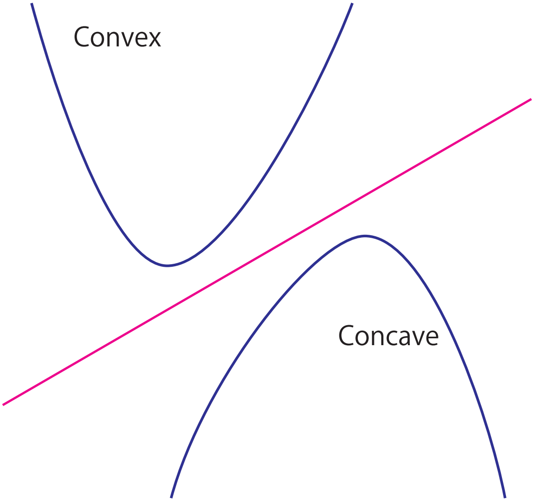
(b) $\exists y \in \mathbb{R}^q$ s.t. $A^\top y \in \mathrm{ri}(\mathrm{dom}(f^*))$ and $-y \in \mathrm{ri}(\mathrm{dom}(g^*))$.

If (a) is satisfied, there exists $y^* \in \mathbb{R}^q$ that attains sup of the RHS.
If (b) is satisfied, there exists $x^* \in \mathbb{R}^p$ that attains inf of the LHS.
Under (a) and (b), x^*, y^* are the optimal solutions of the each side iff

$$A^\top y^* \in \partial f(x^*), \quad Ax^* \in \partial g^*(-y^*).$$

→ **Karush-Kuhn-Tucker condition**.

Equivalence to the separation theorem



Applying Fenchel's duality theorem to RERM

RERM (Regularized Empirical Risk Minimizatino):
Let $\ell_i(z_i^\top x) = \ell(y_i, z_i^\top x)$ where (z_i, y_i) is the input-output pair of the i-th observation.

(Primal)

$$\inf_{x \in \mathbb{R}^p} \Bigg\{ \underbrace{\sum_{i=1}^n \ell_i(z_i^\top x)}_{f(Zx)} \quad + \quad \psi(x) \Bigg\}$$

[Fenchel's duality theorem]

$$\inf_{x \in \mathbb{R}^p} \{ f(Zx) + \psi(x) \} = - \inf_{y \in \mathbb{R}^n} \{ f^*(y) + \psi^*(-Z^\top y) \}$$

(Dual)

$$\sup_{y \in \mathbb{R}^n} \Bigg\{ \sum_{i=1}^n \ell_i^*(y_i) + \psi^*(-Z^\top y) \Bigg\}$$

This fact will be used to derive dual coordinate descent alg.

Outline

1. Introduction

2. Short course to convex analysis
   - Convexity and related concepts
   - Duality
   - Smoothness and strong convexity

Smoothness and strong convexity

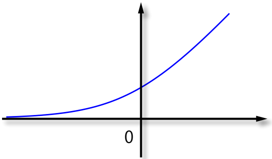
Definition

- Smoothness: the gradient is Lipschitz continuous:

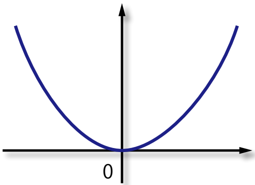
$$\|\nabla f(x) - \nabla f(x')\| \leq L\|x - x'\|.$$

- Strong convexity: $\forall \theta \in (0,1), \forall x, y \in \mathrm{dom}(f)$,

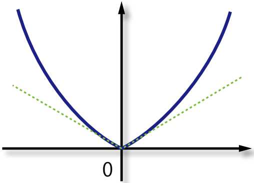
$$\frac{\mu}{2}\theta(1-\theta)\|x-y\|^2 + f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y).$$



Smooth but
not strongly convex



Smooth and
Strongly convex



Strongly convex but
not smooth

Duality between smoothness and strong convexity

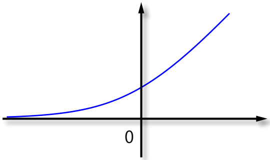
Smoothness and strong convexity is in a relation of duality.

Theorem

Let $f : \mathbb{R}^p \to \bar{\mathbb{R}}$ be proper closed convex.

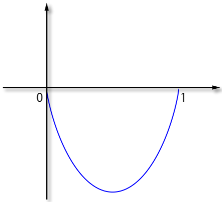
$$f \text{ is } L\text{-smooth} \iff f^* \text{ is } 1/L\text{-strongly convex.}$$

logistic loss



Smooth but
not strongly convex

its dual function



Strongly convex but
not smooth
(gradient $\to \infty$)

T. Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. Annals of Mathematical Statistics, 22: 327–351, 1951.

A. Argyriou, C. A. Micchelli, M. Pontil, and Y. Ying. A spectral regularization framework for multi-task structure learning. In Y. S. J.C. Platt, D. Koller and S. Roweis, editors, Advances in Neural Information Processing Systems 20, pages 25–32, Cambridge, MA, 2008. MIT Press.

O. Banerjee, L. E. Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. Journal of Machine Learning Research, 9:485–516, 2008.

J. Bennett and S. Lanning. The netflix prize. In Proceedings of KDD Cup and Workshop 2007, 2007.

L. Bottou. Online algorithms and stochastic approximations. 1998. URL `http://leon.bottou.org/papers/bottou-98x`. revised, oct 2012.

L. Bottou and Y. LeCun. Large scale online learning. In S. Thrun, L. Saul, and B. Schölkopf, editors, Advances in Neural Information Processing Systems 16. MIT Press, Cambridge, MA, 2004. URL `http://leon.bottou.org/papers/bottou-lecun-2004`.

G. R. Burket. A study of reduced-rank models for multiple prediction, volume 12 of Psychometric monographs. Psychometric Society, 1964.

E. Candès and T. Tao. The power of convex relaxations: Near-optimal matrix completion. IEEE Transactions on Information Theory, 56: 2053–2080, 2009.

E. J. Candès and B. Recht. Exact matrix completion via convex optimization. Foundations of Computational Mathematics, 9(6): 717–772, 2009.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research, 12:2121–2159, 2011.

A. J. Izenman. Reduced-rank regression for the multivariate linear model. Journal of Multivariate Analysis, pages 248–264, 1975.

L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In Proceedings of the 26th International Conference on Machine Learning, 2009.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In C. Burges, L. Bottou, M. Welling,

Z. Ghahramani, and K. Weinberger, editors, Advances in Neural Information Processing Systems 26, pages 315–323. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/4937-accelerating-stochastic-gradient-descent-using-predict pdf.

N. Le Roux, M. Schmidt, and F. R. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, Advances in Neural Information Processing Systems 25, pages 2663–2671. Curran Associates, Inc., 2012. URL http://papers.nips.cc/paper/4633-a-stochastic-gradient-method-with-an-exponential-conve rate-for-finite-training-sets.pdf.

K. Lounici, A. Tsybakov, M. Pontil, and S. van de Geer. Taking advantage of sparsity in multi-task learning. 2009.

N. Meinshausen and P. B uhlmann. High-dimensional graphs and variable selection with the lasso. The Annals of Statistics, 34(3):1436–1462, 2006.

A. Nemirovskii and D. Yudin. On cezari ' s convergence of the steepest descent method for approximating saddle points of convex-concave functions. Soviet Mathematics Doklady, 19(2):576–601, 1978.

A. Nemirovsky and D. Yudin. Problem complexity and method efficiency in optimization. John Wiley, New York, 1983.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. SIAM Journal on Control and Optimization, 30(4): 838–855, 1992.

H. Robbins and S. Monro. A stochastic approximation method. The Annals of Mathematical Statistics, 22(3):400–407, 1951.

F. Rosenblatt. The perceptron: A perceiving and recognizing automaton. Technical Report Technical Report 85-460-1, Project PARA, Cornell Aeronautical Lab., 1957.

D. Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. Journal of Machine Learning Research, 14:567–599, 2013.

Y. Singer and J. C. Duchi. Efficient learning using forward-backward splitting. In Advances in Neural Information Processing Systems, pages 495–503, 2009.

N. Srebro, N. Alon, and T. Jaakkola. Generalization error bounds for collaborative prediction with low-rank matrices. In Advances in Neural Information Processing Systems (NIPS) 17, 2005.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. Journal of Royal Statistical Society: B, 67(1):91–108, 2005.

L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. In Advances in Neural Information Processing Systems 23. 2009.

D. Yogatama and N. A. Smith. Making the most of bag of words: Sentence regularization with alternating direction method of multipliers. In Proceedings of the 31th International Conference on Machine Learning, pages 656–664, 2014.

M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. Biometrika, 94(1):19–35, 2007.