Dual Augmented Lagrangian, Proximal Minimization, and MKL

Ryota Tomioka[1], Taiji Suzuki[1], and Masashi Sugiyama[2]

[1]University of Tokyo
[2]Tokyo Institute of Technology

2009-09-15 @ TU Berlin

Outline

Outline

Outline

Minimum norm reconstruction

- $\boldsymbol{w} \in \mathbb{R}^n$: unknown.
- $y_1, y_2, \ldots, y_m$: observations,
  where $y_i = \langle \boldsymbol{x}_i, \boldsymbol{w} \rangle + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, 1)$.
- Recover $\boldsymbol{w}$ from $\boldsymbol{y}$.

Underdetermined (when $n > m$)

$$\underset{\boldsymbol{w}}{\text{minimize}} \quad \|\boldsymbol{w}\|_0, \quad \text{s.t.} \quad L(\boldsymbol{w}) \leq C,$$

where

$$L(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2,$$

and $\|\boldsymbol{w}\|_0$: the number of non-zero elements in $\boldsymbol{w}$.

Moreover, it is often good to know which features are useful.

However, this is NP hard!

Minimum norm reconstruction

- $\boldsymbol{w} \in \mathbb{R}^n$: unknown.
- $y_1, y_2, \ldots, y_m$: observations,
  where $y_i = \langle \boldsymbol{x}_i, \boldsymbol{w} \rangle + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, 1)$.
- Recover $\boldsymbol{w}$ from $\boldsymbol{y}$.

Underdetermined (when $n > m$)

$$\underset{\boldsymbol{w}}{\text{minimize}} \quad \|\boldsymbol{w}\|_0, \quad \text{s.t.} \quad L(\boldsymbol{w}) \leq C,$$

where

$$L(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2,$$

and $\|\boldsymbol{w}\|_0$: the number of non-zero elements in $\boldsymbol{w}$.

Moreover, it is often good to know which features are useful.

However, this is NP hard!

Minimum norm reconstruction

- $\boldsymbol{w} \in \mathbb{R}^n$: unknown.
- $y_1, y_2, \ldots, y_m$: observations,
  where $y_i = \langle \boldsymbol{x}_i, \boldsymbol{w} \rangle + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, 1)$.
- Recover $\boldsymbol{w}$ from $\boldsymbol{y}$.

Underdetermined (when $n > m$)

$$\underset{\boldsymbol{w}}{\text{minimize}} \quad \|\boldsymbol{w}\|_0, \quad \text{s.t.} \quad L(\boldsymbol{w}) \leq C,$$

where

$$L(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2,$$

and $\|\boldsymbol{w}\|_0$: the number of non-zero elements in $\boldsymbol{w}$.

Moreover, it is often good to know which features are useful.

However, this is NP hard!

Minimum norm reconstruction

- $\boldsymbol{w} \in \mathbb{R}^n$: unknown.
- $y_1, y_2, \ldots, y_m$: observations,
  where $y_i = \langle \boldsymbol{x}_i, \boldsymbol{w} \rangle + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0,1)$.
- Recover $\boldsymbol{w}$ from $\boldsymbol{y}$.

Underdetermined (when $n > m$)

$$\underset{\boldsymbol{w}}{\text{minimize}} \quad \|\boldsymbol{w}\|_0, \quad \text{s.t.} \quad L(\boldsymbol{w}) \leq C,$$

where

$$L(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2,$$

and $\|\boldsymbol{w}\|_0$: the number of non-zero elements in $\boldsymbol{w}$.

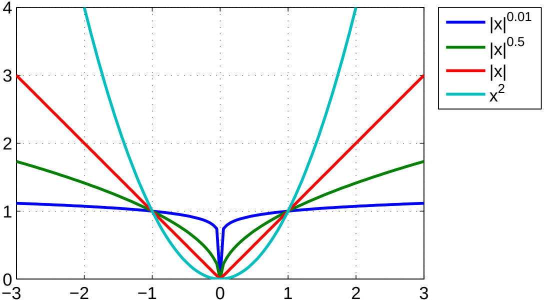
Moreover, it is often good to know which features are useful.

However, this is NP hard!

Convex relaxation

- *p*-norm like functions

$$\|\boldsymbol{w}\|_p^p = \sum_{j=1}^n |w_j|^p : \begin{cases} \text{If } p \geq 1 \text{ convex} \\ \text{If } p < 1 \text{ non-convex} \end{cases}$$

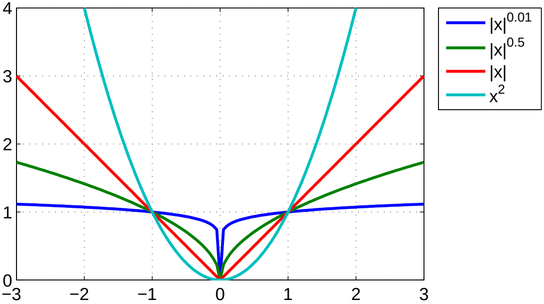


$\|\cdot\|_1$-regularization is the closest to $\|\cdot\|_0$ within convex norm-like functions

Convex relaxation

- *p*-norm like functions

$$\|\boldsymbol{w}\|_p^p = \sum_{j=1}^n |w_j|^p : \begin{cases} \text{If } p \geq 1 \text{ convex} \\ \text{If } p < 1 \text{ non-convex} \end{cases}$$



$\|\cdot\|_1$-regularization is the closest to $\|\cdot\|_0$ within convex norm-like functions

Lasso regression

- Problem 1:

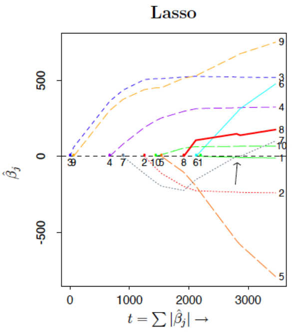
$$\text{minimize} \quad \|\boldsymbol{w}\|_1, \quad \text{s.t. } L(\boldsymbol{w}) \leq C.$$

- Problem 2:

$$\text{minimize} \quad L(\boldsymbol{w}), \quad \text{s.t. } \|\boldsymbol{w}\|_1 \leq C'.$$

- Problem 3:

$$\text{minimize} \quad L(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_1.$$



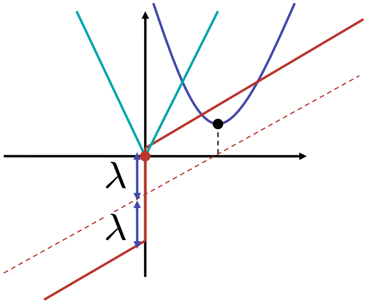
[From Efron et al. (2003)]

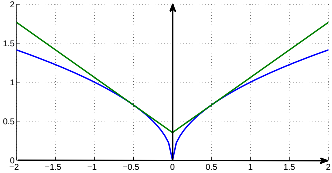
Note:

- Above three problems are equivalent to each other.
- Monotone operation preservs equivalence.
- We focus on the third problem.

Why ℓ_1-regularization?

- The closest to $\|\cdot\|_0$ within convex norm-like functions.
- Non-differentiable at the origin (truncation with finite λ).
- Non-convex regularizers ($p < 1$)
  $\rightarrow$ Iteratively solve (weighted) ℓ_1-regularization.
- Bayesian sparse models (type-II ML)
  $\rightarrow$ Iteratively solve (weighted) ℓ_1-regularization (in special cases).
  (Wipf&Nagarajan, 08)



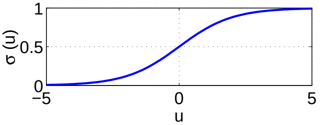


Generalizations

- Generalize the loss term$\cdots$ e.g.,ℓ_1-logistic regression

$$\underset{\boldsymbol{w}\in\mathbb{R}^n}{\text{minimize}} \quad \textstyle\sum_{i=1}^{m} -\log P(y_i|\boldsymbol{x}_i;\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_1$$

$$\text{where} P(y|\boldsymbol{x};\boldsymbol{w}) = \sigma\left(y\left\langle \boldsymbol{w},\boldsymbol{x}\right\rangle\right) \quad (y\in\{-1,+1\})$$



$\left[\sigma(u) = \frac{1}{1+\exp(-u)}\right]$

- Generalize the reg. term$\cdots$ e.g., group lasso (Yuan&Lin,06)

$$\underset{\boldsymbol{w}\in\mathbb{R}^n}{\text{minimize}} \quad L(\boldsymbol{w}) + \lambda\sum_{\mathfrak{g}\in\mathfrak{G}}\|\boldsymbol{w}_{\mathfrak{g}}\|_2$$

where, $\mathfrak{G}$ is a partition of $\{1,\dots,n\}$, $\boldsymbol{w} = \begin{pmatrix} (\boldsymbol{w}_{\mathfrak{g}_1}) \\ (\boldsymbol{w}_{\mathfrak{g}_2}) \\ \vdots \\ (\boldsymbol{w}_{\mathfrak{g}_q}) \end{pmatrix}$, $q = |\mathfrak{G}|$.

Introducing Kernels

Multiple Kernel Learning (MKL) (Lanckriet, Bach, et al., 04)

Let $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_n$ be RKHSs and $K_1, K_2, \ldots, K_n$ be the kernel functions. Use functions $f = \underbrace{f_1}_{\in\mathcal{H}_1} + \underbrace{f_2}_{\in\mathcal{H}_2} + \cdots + \underbrace{f_n}_{\in\mathcal{H}_n}$

$$\underset{f_j\in\mathcal{H}_j, b\in\mathbb{R}}{\text{minimize}} \quad L(f_1 + f_2 + \cdots + f_n + b) + \lambda \sum_{j=1}^{n} \|f_j\|_{\mathcal{H}_j}$$

$\Downarrow$ representer theorem

$$\underset{\alpha_j\in\mathbb{R}^m, b\in\mathbb{R}}{\text{minimize}} \quad f_\ell\Big(\sum_{j=1}^{n} \boldsymbol{K}_j\alpha_j + b\boldsymbol{1}\Big) + \lambda \sum_{j=1}^{n} \|\alpha_j\|_{\boldsymbol{K}_j}$$

where, $\|\alpha_j\|_{\boldsymbol{K}_j} = \sqrt{\alpha_j{}^\top \boldsymbol{K}_j\alpha_j}$.

$\cdots$ nothing but a kernel-weighted group lasso

Introducing Kernels

Multiple Kernel Learning (MKL) (Lanckriet, Bach, et al., 04)
Let $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_n$ be RKHSs and $K_1, K_2, \ldots, K_n$ be the kernel functions. Use functions $f = \underbrace{f_1}_{\in\mathcal{H}_1} + \underbrace{f_2}_{\in\mathcal{H}_2} + \cdots + \underbrace{f_n}_{\in\mathcal{H}_n}$

$$\underset{f_j\in\mathcal{H}_j, b\in\mathbb{R}}{\text{minimize}} \quad L(f_1 + f_2 + \cdots + f_n + b) + \lambda \sum_{j=1}^{n} \|f_j\|_{\mathcal{H}_j}$$

$\Downarrow$ representer theorem

$$\underset{\boldsymbol{\alpha}_j\in\mathbb{R}^m, b\in\mathbb{R}}{\text{minimize}} \quad f_\ell\Big(\sum_{j=1}^{n} \boldsymbol{K}_j\boldsymbol{\alpha}_j + b\mathbf{1}\Big) + \lambda \sum_{j=1}^{n} \|\boldsymbol{\alpha}_j\|_{\boldsymbol{K}_j}$$

where, $\|\boldsymbol{\alpha}_j\|_{\boldsymbol{K}_j} = \sqrt{\boldsymbol{\alpha}_j{}^\top \boldsymbol{K}_j \boldsymbol{\alpha}_j}$.

$\cdots$ nothing but a kernel-weighted group lasso

Modeling assumptions

In many cases the loss term $L(\cdot)$ can be decomposed into a loss function f_ℓ and a design matrix $\boldsymbol{A}$.

- Squared loss

$$f_\ell^{\mathrm{Q}}(\boldsymbol{z}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{z}\|^2, \qquad \boldsymbol{A} = \begin{pmatrix} \boldsymbol{x}_1{}^\top \\ \vdots \\ \boldsymbol{x}_m{}^\top \end{pmatrix}$$

$$\Rightarrow f_\ell^{\mathrm{Q}}(\boldsymbol{A}\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{m}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{w}\rangle)^2$$

- Logistic loss

$$f_\ell^{\mathrm{L}}(\boldsymbol{z}) = \sum_{i=1}^{m}\log(1 + \exp(-y_i z_i)), \qquad \boldsymbol{A} = \begin{pmatrix} \boldsymbol{x}_1{}^\top \\ \vdots \\ \boldsymbol{x}_m{}^\top \end{pmatrix}$$

$$\Rightarrow f_\ell^{\mathrm{L}}(\boldsymbol{A}\boldsymbol{w}) = \sum_{i=1}^{m} -\log\sigma(y_i\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle)$$

Outline

Objective

Develop an optimization algorithm for the problem:

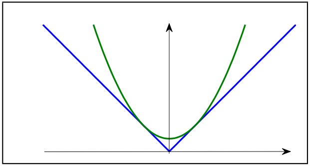
$$\underset{\boldsymbol{w}\in\mathbb{R}^n}{\text{minimize}} \quad f_\ell(\boldsymbol{A}\boldsymbol{w}) + \phi_\lambda(\boldsymbol{w}).$$

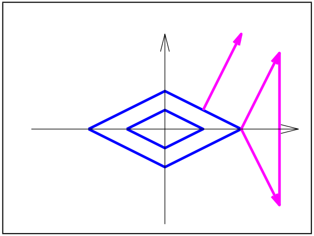
- $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ (m: #observations, n: #unknowns).
- f_ℓ is convex and twice differentiable.
- $\phi_\lambda(\boldsymbol{w})$ is convex but possibly non-differentiable, e.g., $\phi_\lambda(\boldsymbol{w}) = \lambda\|\boldsymbol{w}\|_1$.
- $\eta\phi_\lambda = \phi_{\eta\lambda}$.
- We are interested in algorithms for general f_ℓ ($\leftrightarrow$ LARS).

Where does the difficulty come from?

Conventional view: the non-differentiability of $\phi_\lambda(\boldsymbol{w})$

- Upper bound the regularizer from above with a differentiable function.
  - FOCUSS (Rao & Kreutz-Delgado, 99)
  - Majorization-Minimization (Figueiredo et al., 07)
- Explicitly handle the non-differentiability.
  - Sub-gradient L-BFGS (Andrew & Gao, 07; Yu et al., 08)





Our view: the coupling between variables introduced by $\boldsymbol{A}$.

Where does the difficulty come from?

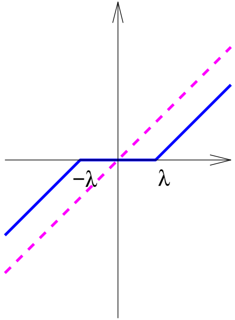
Our view: the coupling between variables introduced by $\boldsymbol{A}$.

In fact, when $\boldsymbol{A} = \boldsymbol{I}_n$

$$\min_{\boldsymbol{w}\in\mathbb{R}^n}\left(\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{w}\|_2^2+\lambda\|\boldsymbol{w}\|_1\right)=\sum_{j=1}^{n}\min_{w_j\in\mathbb{R}}\left(\frac{1}{2}(y_j-w_j)^2+\lambda|w_j|\right).$$

$$\Rightarrow \quad w_j^* = \mathrm{ST}_\lambda(y_j)$$
$$= \begin{cases} y_j-\lambda & (\lambda\le y_j),\\ 0 & (-\lambda\le y_j\le\lambda),\\ y_j+\lambda & (y_j\le-\lambda).\end{cases}$$

min is obtained analytically!



We focus on ϕ_λ for which the above min can be obtained analytically

Earlier study

Iterative Shrinkage/Thresholding (IST) (Figueiredo&Nowak, 03; Daubechies et al., 04;...):

Algorithm

1. Choose an initial solution $\boldsymbol{w}^0$.
2. Repeat until some stopping criterion is satisfied:

$$\boldsymbol{w}^{t+1} \leftarrow \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \left(Q_{\eta_t}(\boldsymbol{w};\boldsymbol{w}^t) + \phi_\lambda(\boldsymbol{w})\right)$$

where

$$Q_\eta(\boldsymbol{w};\boldsymbol{w}^t) = \underbrace{L(\boldsymbol{w}^t) + \nabla L^\top(\boldsymbol{w}^t)(\boldsymbol{w}-\boldsymbol{w}^t)}_{\text{(1) Linearly approximate the loss term.}} + \frac{1}{2\eta}\underbrace{\|\boldsymbol{w}-\boldsymbol{w}^t\|_2^2}_{\text{(2) penalize dist}^2\text{ from the last iterate.}}.$$

Note: minimizing $Q_\eta(\boldsymbol{w};\boldsymbol{w}^t)$ gives the ordinary gradient step.

Earlier study: IST

$$\underset{\boldsymbol{w}\in\mathbb{R}^n}{\mathrm{argmin}}\left(Q_{\eta_t}(\boldsymbol{w};\boldsymbol{w}^t)+\phi_\lambda(\boldsymbol{w})\right)$$

$$=\underset{\boldsymbol{w}\in\mathbb{R}^n}{\mathrm{argmin}}\left(\text{const.}+\nabla L^\top(\boldsymbol{w}^t)(\boldsymbol{w}-\boldsymbol{w}^t)+\frac{1}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{w}^t\|_2^2+\phi_\lambda(\boldsymbol{w})\right)$$

$$=\underbrace{\underset{\boldsymbol{w}\in\mathbb{R}^n}{\mathrm{argmin}}\left(\frac{1}{2\eta_t}\|\boldsymbol{w}-\tilde{\boldsymbol{w}}^t\|_2^2+\phi_\lambda(\boldsymbol{w})\right)}_{\text{assume this min can be obtained analytically}}=:\mathrm{ST}_{\eta_t\lambda}(\tilde{\boldsymbol{w}}^t)$$

where $\tilde{\boldsymbol{w}}^t=\boldsymbol{w}^t-\eta_t\nabla L(\boldsymbol{w}^t)$ (gradient step)

Finally,

$$\boldsymbol{w}^{t+1}\leftarrow\underbrace{\mathrm{ST}_{\eta_t\lambda}}_{\text{shrink}}\underbrace{\left(\boldsymbol{w}^t-\eta_t\nabla L(\boldsymbol{w}^t)\right)}_{\text{gradient step}}$$

- Pro : easy to implement.
- Con : bad for poorly conditioned $\boldsymbol{A}$.

Earlier study: IST

$$\underset{\boldsymbol{w}\in\mathbb{R}^n}{\operatorname{argmin}} \left(Q_{\eta_t}(\boldsymbol{w};\boldsymbol{w}^t) + \phi_\lambda(\boldsymbol{w})\right)$$

$$= \underset{\boldsymbol{w}\in\mathbb{R}^n}{\operatorname{argmin}} \left(\text{const.} + \nabla L^\top(\boldsymbol{w}^t)(\boldsymbol{w}-\boldsymbol{w}^t) + \frac{1}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{w}^t\|_2^2 + \phi_\lambda(\boldsymbol{w})\right)$$

$$= \underbrace{\underset{\boldsymbol{w}\in\mathbb{R}^n}{\operatorname{argmin}} \left(\frac{1}{2\eta_t}\|\boldsymbol{w}-\tilde{\boldsymbol{w}}^t\|_2^2 + \phi_\lambda(\boldsymbol{w})\right)}_{\text{assume this min can be obtained analytically}} =: \mathrm{ST}_{\eta_t\lambda}(\tilde{\boldsymbol{w}}^t)$$

where $\tilde{\boldsymbol{w}}^t = \boldsymbol{w}^t - \eta_t\nabla L(\boldsymbol{w}^t)$ (gradient step)

Finally,

$$\boldsymbol{w}^{t+1} \leftarrow \underbrace{\mathrm{ST}_{\eta_t\lambda}}_{\text{shrink}}\underbrace{\left(\boldsymbol{w}^t - \eta_t\nabla L(\boldsymbol{w}^t)\right)}_{\text{gradient step}}$$

- Pro : easy to implement.
- Con : bad for poorly conditioned $\boldsymbol{A}$.

Earlier study: IST

$$\operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \left(Q_{\eta_t}(\boldsymbol{w};\boldsymbol{w}^t) + \phi_\lambda(\boldsymbol{w})\right)$$

$$= \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \left(\text{const.} + \nabla L^\top(\boldsymbol{w}^t)(\boldsymbol{w}-\boldsymbol{w}^t) + \frac{1}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{w}^t\|_2^2 + \phi_\lambda(\boldsymbol{w})\right)$$

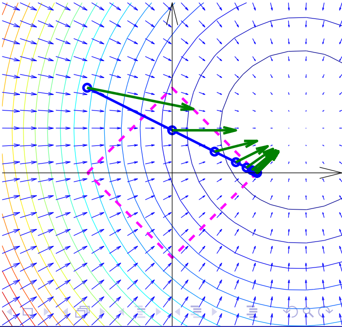
$$= \underbrace{\operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \left(\frac{1}{2\eta_t}\|\boldsymbol{w}-\tilde{\boldsymbol{w}}^t\|_2^2 + \phi_\lambda(\boldsymbol{w})\right)}_{\text{assume this min can be obtained analytically}} =: \mathrm{ST}_{\eta_t\lambda}(\tilde{\boldsymbol{w}}^t)$$

where $\tilde{\boldsymbol{w}}^t = \boldsymbol{w}^t - \eta_t \nabla L(\boldsymbol{w}^t)$ (gradient step)

Finally,

$$\boldsymbol{w}^{t+1} \leftarrow \underbrace{\mathrm{ST}_{\eta_t\lambda}}_{\text{shrink}}\underbrace{\left(\boldsymbol{w}^t - \eta_t\nabla L(\boldsymbol{w}^t)\right)}_{\text{gradient step}}$$

- Pro： easy to implement.
- Con： bad for poorly conditioned $\boldsymbol{A}$.



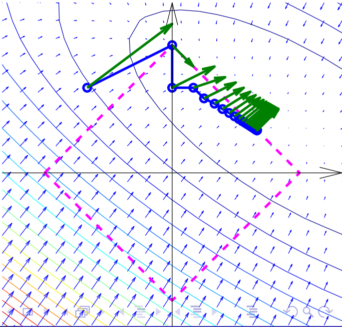
Earlier study: IST

$$\begin{aligned}
&\operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \left(Q_{\eta_t}(\boldsymbol{w};\boldsymbol{w}^t) + \phi_\lambda(\boldsymbol{w})\right)\\
&= \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \left(\text{const.} + \nabla L^\top(\boldsymbol{w}^t)(\boldsymbol{w}-\boldsymbol{w}^t) + \frac{1}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{w}^t\|_2^2 + \phi_\lambda(\boldsymbol{w})\right)\\
&= \underbrace{\operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \left(\frac{1}{2\eta_t}\|\boldsymbol{w}-\tilde{\boldsymbol{w}}^t\|_2^2 + \phi_\lambda(\boldsymbol{w})\right)}_{\text{assume this min can be obtained analytically}} =: \mathrm{ST}_{\eta_t\lambda}(\tilde{\boldsymbol{w}}^t)
\end{aligned}$$

where $\tilde{\boldsymbol{w}}^t = \boldsymbol{w}^t - \eta_t \nabla L(\boldsymbol{w}^t)$ (gradient step)

Finally,

$$\boldsymbol{w}^{t+1} \leftarrow \underbrace{\mathrm{ST}_{\eta_t\lambda}}_{\text{shrink}} \underbrace{\left(\boldsymbol{w}^t - \eta_t \nabla L(\boldsymbol{w}^t)\right)}_{\text{gradient step}}$$



- Pro : easy to implement.
- Con : bad for poorly conditioned $\boldsymbol{A}$.

Summary so far

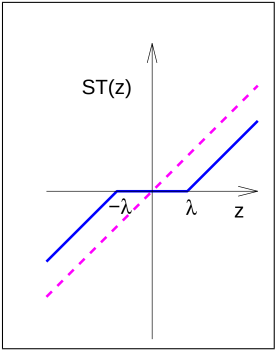
We want to solve:

$$\underset{\boldsymbol{w}\in\mathbb{R}^n}{\text{minimize}} \quad f_\ell(\boldsymbol{A}\boldsymbol{w}) + \phi_\lambda(\boldsymbol{w}).$$

- f_ℓ is convex and twice differentiable.
- $\phi_\lambda(\boldsymbol{w})$ is a convex function for which the minimization:

$$\mathrm{ST}_\lambda(\boldsymbol{z}) = \underset{\boldsymbol{w}\in\mathbb{R}^n}{\text{argmin}} \left(\frac{1}{2}\|\boldsymbol{w}-\boldsymbol{z}\|_2^2 + \phi_\lambda(\boldsymbol{w})\right)$$

  can be carried out analytically, e.g., $\phi_\lambda(\boldsymbol{w}) = \lambda\|\boldsymbol{w}\|_1$.
- Exploit the non-differentiability of ϕ_λ: more sparsity $\rightarrow$ more efficiency.
- Robustify against poor conditioning of $\boldsymbol{A}$.



Outline

Outline

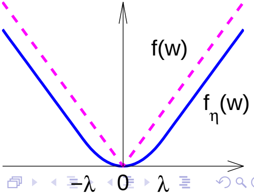
1. Introduction
   - Lasso, group lasso and MKL
   - Objective
2. Method
   - Proximal minimization algorithm
   - Multiple Kernel Learning
3. Experiments
4. Summary

Proximal Minimization (Rockafellar, 1976)

1. Choose an initial solution $\boldsymbol{w}^0$.
2. Repeat until some stopping criterion is satisfied:

$$\boldsymbol{w}^{t+1} \leftarrow \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n}\Big(\underbrace{f_\ell(\boldsymbol{A}\boldsymbol{w})}_{\text{No approximation}} + \phi_\lambda(\boldsymbol{w}) + \frac{1}{2\eta_t} \underbrace{\|\boldsymbol{w}-\boldsymbol{w}^t\|_2^2}_{\text{penalize dist}^2\text{ from the last iterate.}} \Big)$$

- Let $f_\eta(\boldsymbol{w}) = \min_{\tilde{\boldsymbol{w}}\in\mathbb{R}^n}\left(f_\ell(\boldsymbol{A}\tilde{\boldsymbol{w}}) + \phi_\lambda(\tilde{\boldsymbol{w}}) + \frac{1}{2\eta}\|\tilde{\boldsymbol{w}}-\boldsymbol{w}\|_2^2 \right)$.
  - Fact 1: $f_\eta(\boldsymbol{w}) \le f(\boldsymbol{w}) = f_\ell(\boldsymbol{A}\boldsymbol{w}) + \phi_\lambda(\boldsymbol{w})$.
  - Fact 2: $f_\eta(\boldsymbol{w}^*) = f(\boldsymbol{w}^*)$.
- Linearly approximate the loss term → IST



The difference

- IST: linearly approximates the loss term:

$$f_\ell(\boldsymbol{A}\boldsymbol{w}) \simeq f_\ell(\boldsymbol{A}\boldsymbol{w}^t) + (\boldsymbol{w} - \boldsymbol{w}^t)^\top \boldsymbol{A}^\top \nabla f_\ell(\boldsymbol{A}\boldsymbol{w}^t)$$

$\rightarrow$ tightest at the current iterate $\boldsymbol{w}^t$

- DAL (proposed): linearly lower bounds the loss term:

$$f_\ell(\boldsymbol{A}\boldsymbol{w}) = \max_{\boldsymbol{\alpha}\in\mathbb{R}^m} \left(-f_\ell^*(-\boldsymbol{\alpha}) - \boldsymbol{w}^\top \boldsymbol{A}^\top \boldsymbol{\alpha}\right)$$

$\rightarrow$ tightest at the next iterate $\boldsymbol{w}^{t+1}$

The algorithm

IST (Earlier study)

1. Choose an initial solution $\boldsymbol{w}^0$.
2. Repeat until some stopping criterion is satisfied:

$$\boldsymbol{w}^{t+1} \leftarrow \mathrm{ST}_{\eta_t\lambda}\left(\boldsymbol{w}^t + \eta_t \boldsymbol{A}^\top(-\nabla f_\ell(\boldsymbol{A}\boldsymbol{w}^t))\right)$$

Dual Augmented Lagrangian (proposed)

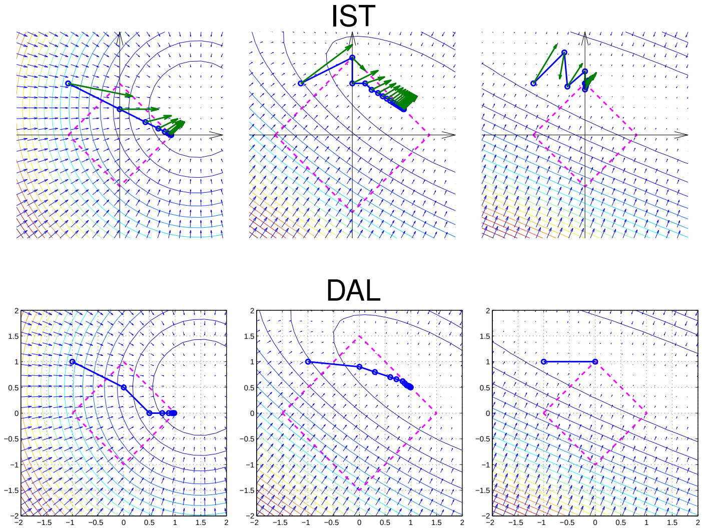
1. Choose an initial solution $\boldsymbol{w}^0$ and a sequence $\eta_0 \leq \eta_1 \leq \cdots$.
2. Repeat until some stopping criterion is satisfied:

$$\boldsymbol{w}^{t+1} \leftarrow \mathrm{ST}_{\eta_t\lambda}\left(\boldsymbol{w}^t + \eta_t \boldsymbol{A}^\top \boldsymbol{\alpha}^t\right)$$

where

$$\boldsymbol{\alpha}^t = \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^m}\left(f_\ell^*(-\boldsymbol{\alpha}) + \frac{1}{2\eta_t}\|\mathrm{ST}_{\eta_t\lambda}(\boldsymbol{w}^t + \eta_t\boldsymbol{A}^\top\boldsymbol{\alpha})\|_2^2\right)$$

Numerical examples



Derivation

$$\boldsymbol{w}^{t+1} \leftarrow \underset{\boldsymbol{w}\in\mathbb{R}^n}{\text{argmin}} \left(f_\ell(\boldsymbol{A}\boldsymbol{w}) + \phi_\lambda(\boldsymbol{w}) + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}^t\|_2^2 \right)$$

$$= \underset{\boldsymbol{w}\in\mathbb{R}^n}{\text{argmin}} \left\{ \max_{\alpha\in\mathbb{R}^m} \left(-f_\ell^*(-\alpha) - \boldsymbol{w}^\top \boldsymbol{A}^\top \alpha \right) + \phi_\lambda(\boldsymbol{w}) + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}^t\|_2^2 \right\}$$

Exchange the order of min and max, calculation, and calculation...

Derivation

$$\boldsymbol{w}^{t+1} \leftarrow \mathop{\mathrm{argmin}}_{\boldsymbol{w}\in\mathbb{R}^n} \left(f_\ell(\boldsymbol{A}\boldsymbol{w}) + \phi_\lambda(\boldsymbol{w}) + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}^t\|_2^2 \right)$$
$$= \mathop{\mathrm{argmin}}_{\boldsymbol{w}\in\mathbb{R}^n} \left\{ \max_{\boldsymbol{\alpha}\in\mathbb{R}^m} \left(-f_\ell^*(-\boldsymbol{\alpha}) - \boldsymbol{w}^\top \boldsymbol{A}^\top \boldsymbol{\alpha} \right) + \phi_\lambda(\boldsymbol{w}) + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}^t\|_2^2 \right\}$$

Exchange the order of min and max, calculation, and calculation...

Augmented Lagrangian

Equality constrained problem:

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad f(\boldsymbol{x}), \qquad \Leftrightarrow \qquad \underset{\boldsymbol{x}}{\text{minimize}}\, L_{hard}(\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}) & (\text{if } \boldsymbol{c}(\boldsymbol{x}) = 0), \\ +\infty & (\text{otherwise}). \end{cases}$$

$$\text{s.t.} \quad \boldsymbol{c}(\boldsymbol{x}) = 0.$$

Ordinary Lagrangian:

$$L(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}) + \boldsymbol{y}^\top \boldsymbol{c}(\boldsymbol{x})$$

Augmented Lagrangian:

$$L_\eta(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}) + \boldsymbol{y}^\top \boldsymbol{c}(\boldsymbol{x}) + \frac{\eta}{2}\|\boldsymbol{c}(\boldsymbol{x})\|_2^2$$

$$L(\boldsymbol{x}, \boldsymbol{y}) \le L_\eta(\boldsymbol{x}, \boldsymbol{y}) \le L_{hard}(\boldsymbol{x})$$

$$\downarrow \min_{\boldsymbol{x}}$$

$$d(\boldsymbol{y}) \le d_\eta(\boldsymbol{y}) \le f(\boldsymbol{x}^*)\text{: primal optimum}$$

Augmented Lagrangian

Equality constrained problem:

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad f(\boldsymbol{x}), \qquad \Leftrightarrow \qquad \underset{\boldsymbol{x}}{\text{minimize}}\, L_{hard}(\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}) & (\text{if } \boldsymbol{c}(\boldsymbol{x}) = 0), \\ +\infty & (\text{otherwise}). \end{cases}$$

$$\text{s.t.} \quad \boldsymbol{c}(\boldsymbol{x}) = 0.$$

Ordinary Lagrangian:

$$L(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}) + \boldsymbol{y}^\top \boldsymbol{c}(\boldsymbol{x})$$

Augmented Lagrangian:

$$L_\eta(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}) + \boldsymbol{y}^\top \boldsymbol{c}(\boldsymbol{x}) + \frac{\eta}{2}\|\boldsymbol{c}(\boldsymbol{x})\|_2^2$$

$$L(\boldsymbol{x}, \boldsymbol{y}) \leq L_\eta(\boldsymbol{x}, \boldsymbol{y}) \leq L_{hard}(\boldsymbol{x})$$

$$\downarrow \min_{\boldsymbol{x}}$$

$$d(\boldsymbol{y}) \leq d_\eta(\boldsymbol{y}) \leq f(\boldsymbol{x}^*)\text{: primal optimum}$$

Augmented Lagrangian

Equality constrained problem:

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad f(\boldsymbol{x}), \qquad \Leftrightarrow \qquad \underset{\boldsymbol{x}}{\text{minimize}}\, L_{hard}(\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}) & (\text{if } \boldsymbol{c}(\boldsymbol{x}) = 0), \\ +\infty & (\text{otherwise}). \end{cases}$$

$$\text{s.t.} \quad \boldsymbol{c}(\boldsymbol{x}) = 0.$$

Ordinary Lagrangian:

$$L(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}) + \boldsymbol{y}^\top \boldsymbol{c}(\boldsymbol{x})$$

Augmented Lagrangian:

$$L_\eta(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}) + \boldsymbol{y}^\top \boldsymbol{c}(\boldsymbol{x}) + \frac{\eta}{2}\|\boldsymbol{c}(\boldsymbol{x})\|_2^2$$

$$L(\boldsymbol{x}, \boldsymbol{y}) \leq L_\eta(\boldsymbol{x}, \boldsymbol{y}) \leq L_{hard}(\boldsymbol{x})$$

$$\downarrow \min_{\boldsymbol{x}}$$

$$d(\boldsymbol{y}) \leq d_\eta(\boldsymbol{y}) \leq f(\boldsymbol{x}^*)\text{: primal optimum}$$

Augmented Lagrangian

Equality constrained problem:

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad f(\boldsymbol{x}), \qquad \Leftrightarrow \qquad \underset{\boldsymbol{x}}{\text{minimize}}\, L_{hard}(\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}) & (\text{if } \boldsymbol{c}(\boldsymbol{x}) = 0), \\ +\infty & (\text{otherwise}). \end{cases}$$

$$\text{s.t.} \quad \boldsymbol{c}(\boldsymbol{x}) = 0.$$

Ordinary Lagrangian:

$$L(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}) + \boldsymbol{y}^\top \boldsymbol{c}(\boldsymbol{x})$$

Augmented Lagrangian:

$$L_\eta(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}) + \boldsymbol{y}^\top \boldsymbol{c}(\boldsymbol{x}) + \frac{\eta}{2}\|\boldsymbol{c}(\boldsymbol{x})\|_2^2$$

$$L(\boldsymbol{x}, \boldsymbol{y}) \le L_\eta(\boldsymbol{x}, \boldsymbol{y}) \le L_{hard}(\boldsymbol{x})$$

$$\downarrow \min_{\boldsymbol{x}}$$

$$d(\boldsymbol{y}) \le d_\eta(\boldsymbol{y}) \le f(\boldsymbol{x}^*)\text{: primal optimum}$$

Augmented Lagrangian

Equality constrained problem:

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad f(\boldsymbol{x}), \qquad \Leftrightarrow \qquad \underset{\boldsymbol{x}}{\text{minimize}}\, L_{hard}(\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}) & (\text{if } \boldsymbol{c}(\boldsymbol{x}) = 0), \\ +\infty & (\text{otherwise}). \end{cases}$$

$$\text{s.t.} \quad \boldsymbol{c}(\boldsymbol{x}) = 0.$$

Ordinary Lagrangian:

$$L(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}) + \boldsymbol{y}^\top \boldsymbol{c}(\boldsymbol{x})$$

Augmented Lagrangian:

$$L_\eta(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}) + \boldsymbol{y}^\top \boldsymbol{c}(\boldsymbol{x}) + \frac{\eta}{2}\|\boldsymbol{c}(\boldsymbol{x})\|_2^2$$

$$L(\boldsymbol{x}, \boldsymbol{y}) \le L_\eta(\boldsymbol{x}, \boldsymbol{y}) \le L_{hard}(\boldsymbol{x})$$

$$\downarrow \min_{\boldsymbol{x}}$$

$$d(\boldsymbol{y}) \le d_\eta(\boldsymbol{y}) \le f(\boldsymbol{x}^*)\text{: primal optimum}$$

Augmented Lagrangian

Equality constrained problem:

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad f(\boldsymbol{x}), \qquad \Leftrightarrow \qquad \underset{\boldsymbol{x}}{\text{minimize}}\, L_{hard}(\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}) & (\text{if } \boldsymbol{c}(\boldsymbol{x}) = 0), \\ +\infty & (\text{otherwise}). \end{cases}$$

$$\text{s.t.} \quad \boldsymbol{c}(\boldsymbol{x}) = 0.$$

Ordinary Lagrangian:

$$L(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}) + \boldsymbol{y}^\top \boldsymbol{c}(\boldsymbol{x})$$

Augmented Lagrangian:

$$L_\eta(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}) + \boldsymbol{y}^\top \boldsymbol{c}(\boldsymbol{x}) + \frac{\eta}{2}\|\boldsymbol{c}(\boldsymbol{x})\|_2^2$$

$$L(\boldsymbol{x}, \boldsymbol{y}) \le L_\eta(\boldsymbol{x}, \boldsymbol{y}) \le L_{hard}(\boldsymbol{x})$$

$$\downarrow \min_{\boldsymbol{x}}$$

$$d(\boldsymbol{y}) \le d_\eta(\boldsymbol{y}) \le f(\boldsymbol{x}^*)\text{: primal optimum}$$

Augmented Lagrangian Algorithm (Hestenes, 69; Powell, 69)

1. Choose an initial multiplier $\boldsymbol{y}^0$ and a sequence $\eta_0 \leq \eta_1 \leq \cdots$.
2. Update the multiplier:

$$\boldsymbol{y}^{t+1} \leftarrow \boldsymbol{y}^t + \eta_t \boldsymbol{c}(\boldsymbol{x}^t)$$

where

$$\boldsymbol{x}^t = \operatorname*{argmin}_{\boldsymbol{x}} L_{\eta_t}(\boldsymbol{x}, \boldsymbol{y}^t)$$

Note

- The multiplier $\boldsymbol{y}^t$ is updated as long as the constraint is violated.
- AL method $\Leftrightarrow$ proximal minimization in the dual (Rockafellar, 76).

$$\boldsymbol{y}^{t+1} \leftarrow \operatorname*{argmax}_{\boldsymbol{y}} \Big(\underbrace{d(\boldsymbol{y})}_{=\min_{\boldsymbol{x}}\left(f(\boldsymbol{x}) + \boldsymbol{y}^\top \boldsymbol{c}(\boldsymbol{x})\right)} - \frac{1}{2\eta_t} \|\boldsymbol{y} - \boldsymbol{y}^t\|_2^2 \Big)$$

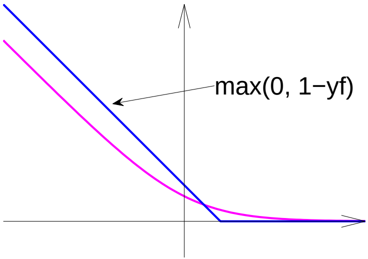
Outline

Objective

Learn a linear combination of kernels from data.

$$\underset{f \in \mathcal{H}, b \in \mathbb{R}, \boldsymbol{d} \in \mathbb{R}^n}{\text{minimize}} \quad \sum_{i=1}^{m} \ell_i(f(x_i) + b) + \frac{\lambda}{2} \|f\|_{\mathcal{H}(d)}^2$$

$$\text{s.t.} \quad \boldsymbol{K}(d) = \sum_{i=1}^{n} d_j \boldsymbol{K}_j, \quad d_j \geq 0, \quad \sum_j d_j \leq 1.$$



Representer theorem

$$\underset{\boldsymbol{\alpha}\in\mathbb{R}^m, b\in\mathbb{R}, \boldsymbol{d}\in\mathbb{R}^n}{\text{minimize}} \quad L(\boldsymbol{K}(d)\boldsymbol{\alpha} + b\mathbf{1}) + \frac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{K}(d)\boldsymbol{\alpha}$$

$$\text{s.t.} \quad \boldsymbol{K}(d) = \sum_{i=1}^{n} d_j \boldsymbol{K}_j, \quad d_j \geq 0, \quad \sum_j d_j \leq 1.$$

Relaxing the regularization term

$$\underset{\boldsymbol{\alpha}\in\mathbb{R}^m, b\in\mathbb{R}, \boldsymbol{d}\in\mathbb{R}^n}{\text{minimize}} \quad L(\boldsymbol{K}(d)\boldsymbol{\alpha} + b\mathbf{1}) + \frac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{K}(d)\boldsymbol{\alpha}$$

$$\text{s.t.} \quad \boldsymbol{K}(d) = \sum_{i=1}^{n} d_j \boldsymbol{K}_j, \quad d_j \geq 0, \quad \sum_j d_j \leq 1.$$

Introduce auxiliary variables α_j ($j = 1, \ldots, n$)

$$\boldsymbol{\alpha}^\top \boldsymbol{K}(d)\boldsymbol{\alpha} = \min_{\alpha_j\in\mathbb{R}^m} \left(\sum_{j=1}^{n} \frac{\alpha_j{}^\top \boldsymbol{K}_j \alpha_j}{d_j} \right) \qquad \text{s.t.} \sum_{j=1}^{n} \boldsymbol{K}_j \alpha_j = \boldsymbol{K}(d)\boldsymbol{\alpha}$$

(Proof) Introduce Lagrangian multiplier β and minimize

$$\frac{1}{2}\sum_{j=1}^{n} \frac{\alpha_j{}^\top \boldsymbol{K}_j \alpha_j}{d_j} + \beta^\top \left(\boldsymbol{K}(d)\alpha - \sum_{j=1}^{n} \boldsymbol{K}_j \alpha_j \right).$$

$\alpha_j = d_j \beta, \quad \beta = \alpha$

Relaxing the regularization term

$$\underset{\boldsymbol{\alpha}\in\mathbb{R}^m, b\in\mathbb{R}, \boldsymbol{d}\in\mathbb{R}^n}{\text{minimize}} \quad L(\boldsymbol{K}(d)\boldsymbol{\alpha} + b\mathbf{1}) + \frac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{K}(d)\boldsymbol{\alpha}$$

$$\text{s.t.} \quad \boldsymbol{K}(d) = \sum_{j=1}^{n} d_j \boldsymbol{K}_j, \quad d_j \geq 0, \quad \sum_j d_j \leq 1.$$

Introduce auxiliary variables $\boldsymbol{\alpha}_j$ $(j = 1, \ldots, n)$

$$\boldsymbol{\alpha}^\top \boldsymbol{K}(d)\boldsymbol{\alpha} = \min_{\boldsymbol{\alpha}_j\in\mathbb{R}^m} \left(\sum_{j=1}^{n} \frac{\boldsymbol{\alpha}_j^\top \boldsymbol{K}_j \boldsymbol{\alpha}_j}{d_j} \right) \qquad \text{s.t.} \sum_{j=1}^{n} \boldsymbol{K}_j \boldsymbol{\alpha}_j = \boldsymbol{K}(d)\boldsymbol{\alpha}$$

(Proof) Introduce Lagrangian multiplier β and minimize

$$\frac{1}{2}\sum_{j=1}^{n} \frac{\boldsymbol{\alpha}_j^\top \boldsymbol{K}_j \boldsymbol{\alpha}_j}{d_j} + \beta^\top \left(\boldsymbol{K}(d)\boldsymbol{\alpha} - \sum_{j=1}^{n} \boldsymbol{K}_j \boldsymbol{\alpha}_j \right).$$

$\boldsymbol{\alpha}_j = d_j \beta, \quad \beta = \boldsymbol{\alpha}$

Relaxing the regularization term

$$\underset{\boldsymbol{\alpha}\in\mathbb{R}^m, b\in\mathbb{R}, \boldsymbol{d}\in\mathbb{R}^n}{\text{minimize}} \quad L(\boldsymbol{K}(d)\boldsymbol{\alpha} + b\mathbf{1}) + \frac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{K}(d)\boldsymbol{\alpha}$$

$$\text{s.t.} \quad \boldsymbol{K}(d) = \sum_{j=1}^n d_j \boldsymbol{K}_j, \quad d_j \geq 0, \quad \sum_j d_j \leq 1.$$

Introduce auxiliary variables $\boldsymbol{\alpha}_j$ $(j = 1, \ldots, n)$

$$\boldsymbol{\alpha}^\top \boldsymbol{K}(d)\boldsymbol{\alpha} = \min_{\boldsymbol{\alpha}_j\in\mathbb{R}^m} \left(\sum_{j=1}^n \frac{\boldsymbol{\alpha}_j{}^\top \boldsymbol{K}_j \boldsymbol{\alpha}_j}{d_j} \right) \qquad \text{s.t.} \sum_{j=1}^n \boldsymbol{K}_j \boldsymbol{\alpha}_j = \boldsymbol{K}(d)\boldsymbol{\alpha}$$

(Proof) Introduce Lagrangian multiplier $\boldsymbol{\beta}$ and minimize

$$\frac{1}{2}\sum_{j=1}^n \frac{\boldsymbol{\alpha}_j{}^\top \boldsymbol{K}_j \boldsymbol{\alpha}_j}{d_j} + \boldsymbol{\beta}^\top \left(\boldsymbol{K}(d)\boldsymbol{\alpha} - \sum_{j=1}^n \boldsymbol{K}_j \boldsymbol{\alpha}_j \right).$$

$\boldsymbol{\alpha}_j = d_j \boldsymbol{\beta}, \;\; \boldsymbol{\beta} = \boldsymbol{\alpha}$

Minimization of the upper-bound

$$\underset{\boldsymbol{\alpha}_j\in\mathbb{R}^m, b\in\mathbb{R}, \boldsymbol{d}\in\mathbb{R}^n}{\text{minimize}} \quad L\Big(\sum_{j=1}^n \boldsymbol{K}_j\boldsymbol{\alpha}_j + b\mathbf{1}\Big) + \frac{\lambda}{2}\sum_{j=1}^n \frac{\boldsymbol{\alpha}_j^\top \boldsymbol{K}_j \boldsymbol{\alpha}_j}{d_j}$$

$$\text{s.t.} \quad d_j \geq 0, \quad \sum_j d_j \leq 1.$$

$$\sum_{j=1}^n \frac{\boldsymbol{\alpha}_j^\top \boldsymbol{K}_j \boldsymbol{\alpha}_j}{d_j} = \sum_{j=1}^n d_j \frac{\boldsymbol{\alpha}_j^\top \boldsymbol{K}_j \boldsymbol{\alpha}_j}{d_j^2} = \sum_{j=1}^n d_j \left(\frac{\|\boldsymbol{\alpha}_j\|_{\boldsymbol{K}_j}}{d_j}\right)^2$$

$$\geq \left(\sum_{j=1}^n d_j \frac{\|\boldsymbol{\alpha}_j\|_{\boldsymbol{K}_j}}{d_j}\right)^2 \quad \left(\begin{array}{c}\sum_j d_j = 1 \\ \text{Jensen's inequality}\end{array}\right)$$

$$= \left(\sum_{j=1}^n \|\boldsymbol{\alpha}_j\|_{\boldsymbol{K}_j}\right)^2$$

↑ linear sum of RKHS norms

Minimization of the upper-bound

$$\underset{\alpha_j \in \mathbb{R}^m, b \in \mathbb{R}, \boldsymbol{d} \in \mathbb{R}^n}{\text{minimize}} \quad L\Big(\sum_{j=1}^{n} \boldsymbol{K}_j \alpha_j + b\mathbf{1}\Big) + \frac{\lambda}{2} \sum_{j=1}^{n} \frac{\alpha_j{}^\top \boldsymbol{K}_j \alpha_j}{d_j}$$

$$\text{s.t.} \quad d_j \geq 0, \quad \sum_j d_j \leq 1.$$

$$\sum_{j=1}^{n} \frac{\alpha_j{}^\top \boldsymbol{K}_j \alpha_j}{d_j} = \sum_{j=1}^{n} d_j \frac{\alpha_j{}^\top \boldsymbol{K}_j \alpha_j}{d_j^2} = \sum_{j=1}^{n} d_j \left(\frac{\|\alpha_j\|_{\boldsymbol{K}_j}}{d_j}\right)^2$$

$$\geq \left(\sum_{j=1}^{n} d_j \frac{\|\alpha_j\|_{\boldsymbol{K}_j}}{d_j}\right)^2 \qquad \left(\begin{array}{c} \sum_j d_j = 1 \\ \text{Jensen's inequality} \end{array}\right)$$

$$= \left(\sum_{j=1}^{n} \|\alpha_j\|_{\boldsymbol{K}_j}\right)^2$$

↑ linear sum of RKHS norms

Minimization of the upper-bound

$$\underset{\boldsymbol{\alpha}_j\in\mathbb{R}^m, b\in\mathbb{R}, \boldsymbol{d}\in\mathbb{R}^n}{\text{minimize}} \quad L\Big(\sum_{j=1}^{n}\boldsymbol{K}_j\boldsymbol{\alpha}_j + b\boldsymbol{1}\Big) + \frac{\lambda}{2}\sum_{j=1}^{n}\frac{\boldsymbol{\alpha}_j^\top\boldsymbol{K}_j\boldsymbol{\alpha}_j}{d_j}$$

$$\text{s.t.} \quad d_j \geq 0, \quad \sum_j d_j \leq 1.$$

$$\sum_{j=1}^{n}\frac{\boldsymbol{\alpha}_j^\top\boldsymbol{K}_j\boldsymbol{\alpha}_j}{d_j} = \sum_{j=1}^{n} d_j \frac{\boldsymbol{\alpha}_j^\top\boldsymbol{K}_j\boldsymbol{\alpha}_j}{d_j^2} = \sum_{j=1}^{n} d_j\left(\frac{\|\boldsymbol{\alpha}_j\|_{\boldsymbol{K}_j}}{d_j}\right)^2$$

$$\geq \left(\sum_{j=1}^{n} d_j \frac{\|\boldsymbol{\alpha}_j\|_{\boldsymbol{K}_j}}{d_j}\right)^2 \qquad \left(\begin{array}{l}\sum_j d_j = 1 \\ \text{Jensen's inequality}\end{array}\right)$$

$$= \left(\sum_{j=1}^{n}\|\boldsymbol{\alpha}_j\|_{\boldsymbol{K}_j}\right)^2$$

↑ linear sum of RKHS norms

Minimization of the upper-bound

$$\underset{\boldsymbol{\alpha}_j\in\mathbb{R}^m, b\in\mathbb{R}, \boldsymbol{d}\in\mathbb{R}^n}{\text{minimize}} \quad L\Big(\sum_{j=1}^{n}\boldsymbol{K}_j\boldsymbol{\alpha}_j + b\boldsymbol{1}\Big) + \frac{\lambda}{2}\sum_{j=1}^{n}\frac{\boldsymbol{\alpha}_j{}^\top\boldsymbol{K}_j\boldsymbol{\alpha}_j}{d_j}$$

$$\text{s.t.} \quad d_j \geq 0, \quad \sum_j d_j \leq 1.$$

$$\sum_{j=1}^{n}\frac{\boldsymbol{\alpha}_j{}^\top\boldsymbol{K}_j\boldsymbol{\alpha}_j}{d_j} = \sum_{j=1}^{n} d_j \frac{\boldsymbol{\alpha}_j{}^\top\boldsymbol{K}_j\boldsymbol{\alpha}_j}{d_j^2} = \sum_{j=1}^{n} d_j \left(\frac{\|\boldsymbol{\alpha}_j\|_{\boldsymbol{K}_j}}{d_j}\right)^2$$

$$\geq \left(\sum_{j=1}^{n} d_j \frac{\|\boldsymbol{\alpha}_j\|_{\boldsymbol{K}_j}}{d_j}\right)^2 \quad \left(\begin{array}{l}\sum_j d_j = 1 \\ \text{Jensen's inequality}\end{array}\right)$$

$$= \left(\sum_{j=1}^{n}\|\boldsymbol{\alpha}_j\|_{\boldsymbol{K}_j}\right)^2$$

↑ linear sum of RKHS norms

Minimization of the upper-bound

$$\underset{\boldsymbol{\alpha}_j\in\mathbb{R}^m, b\in\mathbb{R}, \boldsymbol{d}\in\mathbb{R}^n}{\text{minimize}} \quad L\left(\sum_{j=1}^{n}\boldsymbol{K}_j\boldsymbol{\alpha}_j + b\mathbf{1}\right) + \frac{\lambda}{2}\sum_{j=1}^{n}\frac{\boldsymbol{\alpha}_j{}^\top\boldsymbol{K}_j\boldsymbol{\alpha}_j}{d_j}$$

$$\text{s.t.} \quad d_j \geq 0, \quad \sum_j d_j \leq 1.$$

$$\sum_{j=1}^{n}\frac{\boldsymbol{\alpha}_j{}^\top\boldsymbol{K}_j\boldsymbol{\alpha}_j}{d_j} = \sum_{j=1}^{n} d_j \frac{\boldsymbol{\alpha}_j{}^\top\boldsymbol{K}_j\boldsymbol{\alpha}_j}{d_j^2} = \sum_{j=1}^{n} d_j \left(\frac{\|\boldsymbol{\alpha}_j\|_{\boldsymbol{K}_j}}{d_j}\right)^2$$

$$\geq \left(\sum_{j=1}^{n} d_j \frac{\|\boldsymbol{\alpha}_j\|_{\boldsymbol{K}_j}}{d_j}\right)^2 \qquad \left(\begin{array}{l}\sum_j d_j = 1 \\ \text{Jensen's inequality}\end{array}\right)$$

$$= \left(\sum_{j=1}^{n}\|\boldsymbol{\alpha}_j\|_{\boldsymbol{K}_j}\right)^2$$

↑ linear sum of RKHS norms

Minimization of the upper-bound

$$\underset{\boldsymbol{\alpha}_j\in\mathbb{R}^m, b\in\mathbb{R}, \boldsymbol{d}\in\mathbb{R}^n}{\text{minimize}} \quad L\Big(\sum_{j=1}^{n} \boldsymbol{K}_j\boldsymbol{\alpha}_j + b\mathbf{1}\Big) + \frac{\lambda}{2}\sum_{j=1}^{n}\frac{\boldsymbol{\alpha}_j{}^\top\boldsymbol{K}_j\boldsymbol{\alpha}_j}{d_j}$$

$$\text{s.t.} \quad d_j \geq 0, \quad \sum_j d_j \leq 1.$$

$$\sum_{j=1}^{n}\frac{\boldsymbol{\alpha}_j{}^\top\boldsymbol{K}_j\boldsymbol{\alpha}_j}{d_j} = \sum_{j=1}^{n} d_j\frac{\boldsymbol{\alpha}_j{}^\top\boldsymbol{K}_j\boldsymbol{\alpha}_j}{d_j^2} = \sum_{j=1}^{n} d_j\left(\frac{\|\boldsymbol{\alpha}_j\|_{\boldsymbol{K}_j}}{d_j}\right)^2$$

$$\geq \left(\sum_{j=1}^{n} d_j\frac{\|\boldsymbol{\alpha}_j\|_{\boldsymbol{K}_j}}{d_j}\right)^2 \qquad \left(\begin{array}{l}\sum_j d_j = 1\\ \text{Jensen's inequality}\end{array}\right)$$

$$= \left(\sum_{j=1}^{n}\|\boldsymbol{\alpha}_j\|_{\boldsymbol{K}_j}\right)^2$$

↑ linear sum of RKHS norms

Equivalence of the two formulations

Penalizing the square of linear sum of RKHS norms (Bach et al.)

$$\underset{\boldsymbol{\alpha}_j \in \mathbb{R}^m, b \in \mathbb{R}}{\text{minimize}} \quad L\Big(\sum_{j=1}^{n} \boldsymbol{K}_j \boldsymbol{\alpha}_j + b\mathbf{1}\Big) + \frac{\lambda}{2}\Big(\sum_{j=1}^{n} \|\boldsymbol{\alpha}_j\|_{\boldsymbol{K}_j}\Big)^2 \qquad \text{(A)}$$

Penalizing of the linear sum of RKHS norms (proposed)

$$\Leftrightarrow \underset{\boldsymbol{\alpha}_j \in \mathbb{R}^m, b \in \mathbb{R}}{\text{minimize}} \quad L\Big(\sum_{j=1}^{n} \boldsymbol{K}_j \boldsymbol{\alpha}_j + b\mathbf{1}\Big) + \lambda' \sum_{j=1}^{n} \|\boldsymbol{\alpha}_j\|_{\boldsymbol{K}_j} \qquad \text{(B)}$$

Optimality of (A):
$$\nabla_{\boldsymbol{\alpha}_j} L + \lambda\Big(\sum_{j=1}^{n} \|\boldsymbol{\alpha}_j\|_{\boldsymbol{K}_j}\Big) \partial_{\boldsymbol{\alpha}_j} \|\boldsymbol{\alpha}_j\|_{\boldsymbol{K}_j} \ni 0$$

Optimality of (B):
$$\nabla_{\boldsymbol{\alpha}_j} L + \lambda' \partial_{\boldsymbol{\alpha}_j} \|\boldsymbol{\alpha}_j\|_{\boldsymbol{K}_j} \ni 0$$

SpicyMKL

DAL + MKL = SpicyMKL (Sparse Iterative MKL)

- The bias term b, and the hinge-loss need spacial care.
- Soft-thresholding per kernel ($\leftrightarrow$ per variable)

$$\mathrm{ST}_\lambda(\boldsymbol{\alpha}_j) = \begin{cases} 0 & (\|\boldsymbol{\alpha}_j\|_{\boldsymbol{K}_j} \leq \lambda) \\ \left(\|\boldsymbol{\alpha}_j\|_{\boldsymbol{K}_j} - \lambda\right) \frac{\boldsymbol{\alpha}_j}{\|\boldsymbol{\alpha}_j\|_{\boldsymbol{K}_j}} & \text{(otherwise)} \end{cases}$$

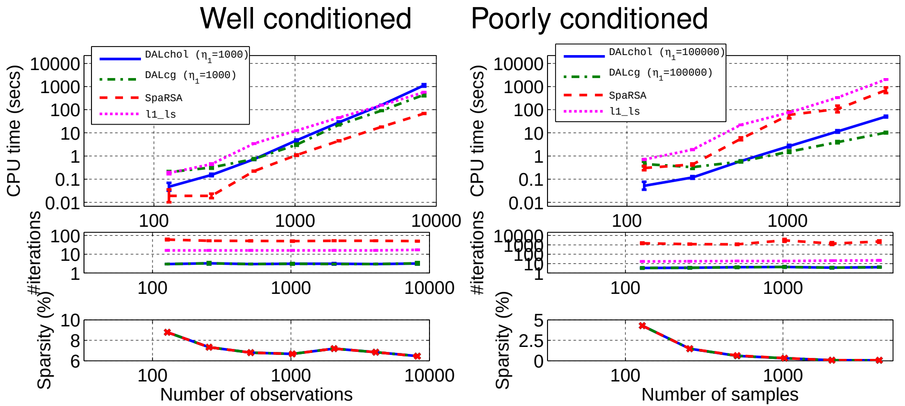
Outline

Experimental setting

- Problem: lasso (square loss + L1 regularization)
- Comparison with:
  - l1_ls (interior-point method)
  - SpaRSA (step-size improved IST)

  (problem specific methods (e.g., LARS) are not considered.)
- Random design matrix $A \in \mathbb{R}^{m \times n}$ (m: #observations, n: #unknowns) generated as:
  - `A=randn(m,n);` (well conditioned)
  - `A=U*diag(1./(1:m))*V';` (poorly conditioned)
- Two settings:
  - Medium Scale ($n = 4m$, $n < 10000$)
  - Large Scale ($m = 1024$, $n < 1e + 6$)

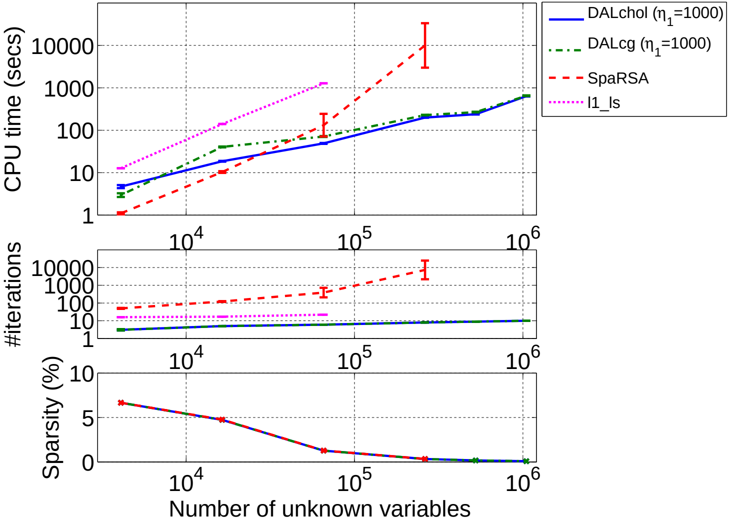
Results (medium scale)



(a) Normal conditioning

(b) Poor conditioning

Results (large scale)



L1-logistic regression

- $m = 1,024$.
- $n = 4,096–32,768$.

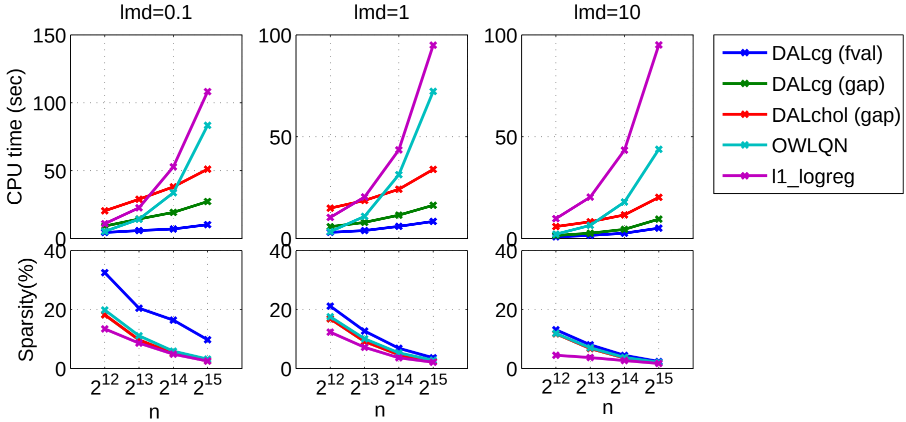


Image classification

- Picked five classes anchor, ant, cannon, chair, cup from Caltech 101 dataset (Fei-Fei et al., 2004).
- Ten 2-class classification problems.
- # kernels 1,760 = Feature extraction (4) $\times$ Spatial subdivision (22) $\times$ Kernel functions (20)
  - Feature extraction: (a) hsvsift, (b) sift (scale auto), (c) sift (scale 4px fixed), (d) sift (scale 8px fixed) (used van de Sande's code)
  - Spatial subdivision and integration: (a) whole image, (b) 2x2 grid, and (c) 4x4 grid + spatial pyramid kernel (Lazebnik et al., 06).
  - Kernel functions: Gaussian RBF kernel and χ^2 kernels using 10 different scale parameters each.

(cf. Gehler & Nowozin, 09)

Outline

Summary

DAL (dual augmented Lagrangian)

- is a dual method in the dual (=primal proximal minimization)
- is efficient when $m \ll n$.
- tolerates poorly conditioned design matrix $\boldsymbol{A}$ better.
- exploits sparsity in the solution (not in the design).
- Legendre transform: linear lower bound instead of linear approximation.

Tasks:

- theoretical analysis.
- cool application.