# Stochastic Optimization
# Part III: Advanced topics of stochastic optimization

[†‡] Taiji Suzuki

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

MLSS2015@Kyoto

# Outline

1. Stochastic optimization for structured regularization
   - Structured regularization
   - Alternating Direction Method of Multipliers (ADMM)
   - Stochastic ADMM for online data
   - Stochastic ADMM for batch data
2. Parallel and distributed optimization
3. Further interesting topics

# Outline

1. Stochastic optimization for structured regularization
   - Structured regularization
   - Alternating Direction Method of Multipliers (ADMM)
   - Stochastic ADMM for online data
   - Stochastic ADMM for batch data
2. Parallel and distributed optimization
3. Further interesting topics

# Outline


1. Stochastic optimization for structured regularization
   - Structured regularization
   - Alternating Direction Method of Multipliers (ADMM)
   - Stochastic ADMM for online data
   - Stochastic ADMM for batch data
2. Parallel and distributed optimization
3. Further interesting topics

# Regularized learning problem

Lasso:

$$\min_{x \in \mathbb{R}^p} \quad \frac{1}{n}\sum_{i=1}^{n}(y_i - z_i^\top x)^2 + \underbrace{\|x\|_1}_{\text{regularization}} \quad .$$

# Regularized learning problem

Lasso:

$$\min_{x \in \mathbb{R}^p} \quad \frac{1}{n}\sum_{i=1}^{n}(y_i - z_i^\top x)^2 + \underbrace{\|x\|_1}_{\text{regularization}} \quad .$$

General regularized learning problem:

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

**Difficulty**: Sparsity inducing regularization is usually **non-smooth**.

# Proximal mapping

Regularized learning problem:

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

$g_t \in \partial_x \left( \frac{1}{n} \sum_{i=1}^n f_i(z_i^\top x) \right) |_{x=x^{(t-1)}}, \; \bar{g}_t = \frac{1}{t} \sum_{\tau=1}^t g_\tau.$
Proximal gradient descent:

$$x^{(t)} = \mathop{\arg\min}_{x \in \mathbb{R}^p} \left\{ g_t^\top x + \tilde{\psi}(x) + \frac{1}{2\eta_t} \|x - x^{(t-1)}\|^2 \right\}.$$

Regularized dual averaging:

$$x^{(t)} = \mathop{\arg\min}_{x \in \mathbb{R}^p} \left\{ \bar{g}_t^\top x + \tilde{\psi}(x) + \frac{1}{2\eta_t} \|x\|^2 \right\}.$$

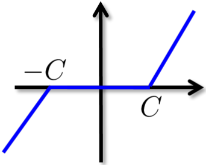
A key computation is the *proximal mapping*:

$$\boxed{\mathrm{prox}(q|\tilde{\psi}) := \mathop{\arg\min}_{x} \left\{ \tilde{\psi}(x) + \frac{1}{2} \|x - q\|^2 \right\}}.$$

# Example of Proximal mapping: $\ell_1$ regularization

$$
\begin{aligned}
\mathrm{prox}(q|C\|\cdot\|_1) &= \arg\min_x \left\{ C\|x\|_1 + \frac{1}{2}\|x-q\|^2 \right\} \\
&= (\mathrm{sign}(q_j)\max(|q_j| - C, 0))_j.
\end{aligned}
$$

→ Soft-thresholding function. Analytic form.




There are also many regularization functions for which computing the prox mapping is difficult.

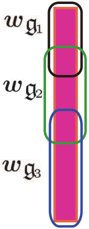
→ **Structured regularization.**

# Examples of structured regularization

- Overlapped group lasso

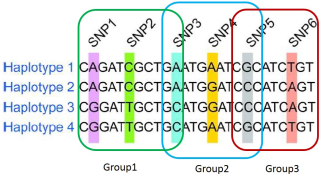
$$\tilde{\psi}(w) = C \sum_{\mathfrak{g} \in \mathfrak{G}} \|w_{\mathfrak{g}}\|_q$$

$(q > 1; typically q = 2, \infty)$




- The groups may have overlap.
- It is difficult to compute the proximal mapping.

Application (1)




Genome Wide Association Study (GWAS)

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

# Application of group reg. (2)

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

The words occurred in the same sentence is grouped:

$$\tilde{\psi}(w) = \sum_{d=1}^{D} \sum_{s=1}^{S_d} \lambda_{d,s} \| w_{(d,s)} \|_2,$$

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

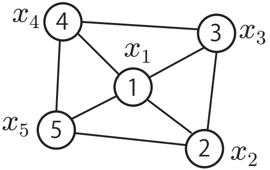
*Table 4.* A review from Amazon dvd review dataset categorized as a positive review. Each line is a sentence identified by the sentence segmenter. There are five sentences in this article. Selected sentences in the learner's copy variables are highlighted in **blue and bold**. We also display the color-coded 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 of romance ( love at first sight , great passion , etc . ) and gangster flicks ( brutality , dagerous machinations , the mysterious don , etc. ) , but it is all done with the crudest humor .** | | (0.42)<br>(0.22)<br>(0.07)<br>(0.48) |
| it ' s the kind of thing you either like viserally and immediately " get " or you don ' t . | | (0.01)<br>(0.01) |
| that is a matter of taste and expectations . | | (0.01) |
| **i enjoyed it and it took me back to the mid80s , when nicolson and turner were in their primes .** | | (0.02)<br>(0.01) |
| the acting is very good , if a bit obviously tongue - in - cheek . | | (0.01) |

# (Generalized) Fused Lasso and TV-denoising

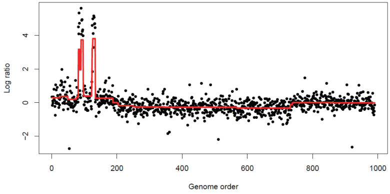
Fused lasso (Tibshirani et al. (2005), Jacob et al. (2009)):

$$\psi(x) = C \sum_{(i,j)\in E} |x_i - x_j|.$$




TV-denoising (Rudin et al., 1992, Chambolle, 2004):

$$\psi(X) = C \sum_{i,j} \sqrt{|X_{i+1,j} - X_{i,j}|^2 + |X_{i,j+1} - X_{i,j}|^2}.$$




Genome data analysis by Fused lasso (Tibshirani and Taylor, 2011)


Image restoration (Mairal et al., 2009)

# 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).

# Solutions

- Developing a sophisticated method for each regularization: Jacob et al. (2009), Yuan et al. (2011).
- Submodular optimization: Bach (2010), Kawahara et al. (2009), Bach et al. (2012)
- Decomposing the proximal mapping: Yu (2013)
- Applying linear transformation to make the regularization simpler.
  $\rightarrow$ **Alternating Direction Method of Multipliers, ADMM**.

# Outline

1. Stochastic optimization for structured regularization
   - Structured regularization
   - Alternating Direction Method of Multipliers (ADMM)
   - Stochastic ADMM for online data
   - Stochastic ADMM for batch data
2. Parallel and distributed optimization
3. Further interesting topics

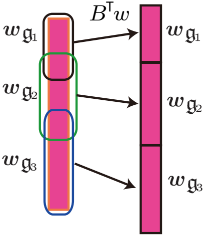
# Linear transformation, Decomposition technique

- Overlapped group lasso $\tilde{\psi}(w) = C \sum_{\mathfrak{g} \in \mathfrak{G}} \|w_{\mathfrak{g}}\|$

※ It is difficult to compute the proximal mapping.

**Solution:**

- Prepare $\psi$ for which proximal mapping is easily computable.
- Let $\psi(B^\top w) = \tilde{\psi}(w)$, and utilize the proximal mapping w.r.t. $\underline{\psi}$.




Decompose into independent groups:

$B^\top w =$

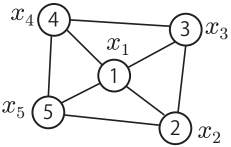

$$\psi(y) = C \sum_{\mathfrak{g}' \in \mathfrak{G}'} \|y_{\mathfrak{g}'}\|$$

$$\mathrm{prox}(q|\psi) = \left( q_{\mathfrak{g}'} \max\left\{ 1 - \frac{C}{\|q_{\mathfrak{g}'}\|}, 0 \right\} \right)_{\mathfrak{g}' \in \mathfrak{G}'}$$

# Another example

- Graph guided regularization

$$\tilde{\psi}(w) = C \sum_{(i,j)\in E} |w_i - w_j|.$$




---

$$\psi(y) = C \sum_{e\in E} |y_e|, \quad y = B^\top w = (w_i - w_j)_{(i,j)\in E}$$

$$\Rightarrow \begin{cases} \psi(B^\top w) = \tilde{\psi}(w), \\ \mathrm{prox}(q|\psi) = \left(q_e \max\left\{1 - \frac{C}{|q_e|}, 0\right\}\right)_{e\in E}. \end{cases}$$

Soft-Thresholding function.

# Optimizing composite objective function with linear constraint

$$\min_x \frac{1}{n}\sum_{i=1}^{n} f_i(z_i^\top x) + \psi(B^\top x)$$

$$\Leftrightarrow \quad \min_{x,y} \frac{1}{n}\sum_{i=1}^{n} f_i(z_i^\top x) + \psi(y) \ \text{ s.t. } \ y = B^\top x.$$

**Augmented Lagrangian**

$$\mathcal{L}(x,y,\lambda) = \frac{1}{n}\sum_i f_i(z_i^\top x) + \psi(y) + \lambda^\top(y - B^\top x) + \frac{\rho}{2}\|y - B^\top x\|^2$$

$$\inf_{x,y}\sup_\lambda \mathcal{L}(x,y,\lambda)$$

yields the optimization of the original problem.

- The augmented Lagrangian is the basis of the method of multipliers (Hestenes, 1969, Powell, 1969, Rockafellar, 1976).

# Method of multipliers

$$\min_{x,y} \; \{\; f(x) + \psi(y) \quad \text{s.t.} \quad Ax + By = 0\}$$

$$\mathcal{L}(w, y, \lambda) = f(x) + \psi(y) + \lambda^\top (Ax + By) + \frac{\rho}{2}\|Ax + By\|^2$$

**Method of multipliers** (Hestenes, 1969, Powell, 1969)

$$(x^{(t)}, y^{(t)}) = \operatorname*{argmin}_{(x,y)} \mathcal{L}(x, y, \lambda^{(t-1)})$$

$$\lambda^{(t)} = \lambda^{(t-1)} + \rho(Ax^{(t)} + By^{(t)})$$

# Method of multipliers

$$\min_{x,y} \ \{\ f(x) + \psi(y) \quad \text{s.t.} \quad Ax + By = 0\}$$

$$\mathcal{L}(w, y, \lambda) = f(x) + \psi(y) + \lambda^\top (Ax + By) + \frac{\rho}{2}\|Ax + By\|^2$$

**Method of multipliers** (Hestenes, 1969, Powell, 1969)

$$(x^{(t)}, y^{(t)}) = \operatorname*{argmin}_{(x,y)} \mathcal{L}(x, y, \lambda^{(t-1)})$$

$$\lambda^{(t)} = \lambda^{(t-1)} + \rho(Ax^{(t)} + By^{(t)})$$

**Remark:**

- Update of $\lambda$ corresponds to gradient ascent of $\mathcal{L}$.
- It is easy to check that

$$\nabla_x (f(x) + \langle \lambda^{(t)}, Ax + By^{(t)} \rangle)|_{x = x^{(t)}} = 0$$

$$\nabla_y (\psi(y) + \langle \lambda^{(t)}, Ax^{(t)} + By \rangle)|_{y = y^{(t)}} = 0.$$

If $Ax^{(t)} + By^{(t)} = 0$, this give the optimality condition.

# Alternating Direction Method of multipliers

**(Douglas-Rachford splitting)**

$$\min_{x,y} \; \{ \, f(x) + \psi(y) \quad \text{s.t.} \quad Ax + By = 0\}$$

$$\mathcal{L}(w, y, \lambda) = f(x) + \psi(y) + \lambda^\top (Ax + By) + \frac{\rho}{2}\|Ax + By\|^2$$

**Alternating Direction Method of Multipliers** (Gabay and Mercier, 1976)

$$x^{(t)} = \operatorname*{argmin}_x \mathcal{L}(x, y^{(t-1)}, \lambda^{(t-1)})$$

$$y^{(t)} = \operatorname*{argmin}_y \mathcal{L}(x^{(t)}, y, \lambda^{(t-1)})$$

$$\lambda^{(t)} = \lambda^{(t-1)} + \rho(Ax^{(t)} + By^{(t)})$$

# Alternating Direction Method of multipliers

**(Douglas-Rachford splitting)**

$$\min_{x,y} \ \{ \ f(x) + \psi(y) \quad \text{s.t.} \quad Ax + By = 0 \}$$

$$\mathcal{L}(w, y, \lambda) = f(x) + \psi(y) + \lambda^\top (Ax + By) + \frac{\rho}{2}\|Ax + By\|^2$$

**Alternating Direction Method of Multipliers** (Gabay and Mercier, 1976)

$$x^{(t)} = \operatorname*{argmin}_x \mathcal{L}(x, y^{(t-1)}, \lambda^{(t-1)})$$

$$y^{(t)} = \operatorname*{argmin}_y \mathcal{L}(x^{(t)}, y, \lambda^{(t-1)})$$

$$\lambda^{(t)} = \lambda^{(t-1)} + \rho(Ax^{(t)} + By^{(t)})$$

- ADMM converges to the optimal (Mota et al., 2011)
- $O(1/t)$ convergence in general (He and Yuan, 2012)
- Linear convergence for a strongly convex objective (Deng and Yin, 2012, Hong and Luo, 2012)

# ADMM for structured regularization

$$\min_x \{f(x) + \psi(B^\top w)\} \Leftrightarrow \min_{x,y} \{f(x) + \psi(y) \text{ s.t. } y = B^\top x\}$$

$$\mathcal{L}(x, y, \lambda) = f(x) + \psi(y) + \lambda^\top (y - B^\top x) + \frac{\rho}{2} \|y - B^\top x\|^2$$

where $f(x) = \frac{1}{n} \sum f_i(z_i^\top x)$

**ADMM for structured regularization**

$$x^{(t)} = \arg\min_x \{f(x) + \lambda^{(t-1)^\top} (-B^\top x) + \frac{\rho}{2} \|y^{(t-1)} - B^\top x\|^2\}$$

$$y^{(t)} = \arg\min_y \{\psi(y) + \lambda^{(t)^\top} y + \frac{\rho}{2} \|y - B^\top x^{(t)}\|^2\}$$

$$(= \mathrm{prox}(B^\top x^{(t)} - \lambda^{(t)}/\rho | \psi/\rho))$$

$$\lambda^{(t)} = \lambda^{(t-1)} - \rho(B^\top x^{(t)} - y^{(t)})$$

- The update of $y$ is given by the proximal mapping w.r.t. simple $\psi$.
  $\rightarrow$ Usually analytic form.

# ADMM for structured regularization

$$\min_x \{f(x) + \psi(B^\top w)\} \Leftrightarrow \min_{x,y} \{f(x) + \psi(y) \text{ s.t. } y = B^\top x\}$$

$$\mathcal{L}(x, y, \lambda) = f(x) + \psi(y) + \lambda^\top (y - B^\top x) + \frac{\rho}{2}\|y - B^\top x\|^2$$

where $f(x) = \frac{1}{n}\sum f_i(z_i^\top x)$

**ADMM for structured regularization**

$$x^{(t)} = \arg\min_x \{f(x) + \lambda^{(t-1)^\top}(-B^\top x) + \frac{\rho}{2}\|y^{(t-1)} - B^\top x\|^2\}$$

$$y^{(t)} = \arg\min_y \{\psi(y) + \lambda^{(t)^\top} y + \frac{\rho}{2}\|y - B^\top x^{(t)}\|^2\}$$

$$(= \mathrm{prox}(B^\top x^{(t)} - \lambda^{(t)}/\rho | \psi/\rho))$$

$$\lambda^{(t)} = \lambda^{(t-1)} - \rho(B^\top x^{(t)} - y^{(t)})$$

- The update of $y$ is given by the proximal mapping w.r.t. simple $\psi$.
  $\rightarrow$ Usually analytic form.
- However, the computation of $\frac{1}{n}\sum_{i=1}^n f_i(z_i^\top w)$ is still heavy.
  $\rightarrow$ **Stochastic version** of ADMM has been developed (Suzuki, 2013, Ouyang et al., 2013, Suzuki, 2014).

# Example: ADMM for Lasso

$$\min_x \left\{ \frac{1}{2n} \|Zx - Y\|^2 + C\|x\|_1 \right\}$$

## ADMM for Lasso

$$x^{(t)} = \left( \frac{Z^\top Z}{2n} + \frac{\rho I}{2} \right)^{-1} \left( \frac{Y}{n} + \lambda^{(t)} - \rho y^{(t-1)} \right)$$

$$y^{(t)} = \mathrm{ST}_{\frac{C}{\rho}}(x^{(t)} - \lambda^{(t)}/\rho)$$

$$\lambda^{(t)} = \lambda^{(t-1)} - \rho(x^{(t)} - y^{(t)})$$

$$\mathrm{ST}_\eta(x) = (\mathrm{sign}(x_j) \max\{|x_j| - \eta, 0\})_j$$

# Outline

1. Stochastic optimization for structured regularization
   - Structured regularization
   - Alternating Direction Method of Multipliers (ADMM)
   - Stochastic ADMM for online data
   - Stochastic ADMM for batch data
2. Parallel and distributed optimization
3. Further interesting topics

# Table of literature

Stochastic methods for regularized learning problems.

| | Normal | ADMM |
|---|---|---|
| Online | • Proximal gradient type (Nesterov, 2007)<br>FOBOS (Duchi and Singer, 2009)<br><br>• Dual averaging type (Nesterov, 2009)<br>RDA (Xiao, 2009) | Online-ADMM (Wang and Banerjee, 2012)<br>SGD-ADMM (Suzuki, 2013, Ouyang et al., 2013)<br>RDA-ADMM (Suzuki, 2013) |
| Batch | SDCA (Shalev-Shwartz and Zhang, 2013)<br>(Stochastic Dual Coordinate Ascent)<br>SAG (Le Roux et al., 2013)<br>(Stochastic Averaging Gradient)<br>SVRG (Johnson and Zhang, 2013)<br>(Stochastic Variance Reduced Gradient) | SDCA-ADMM (Suzuki, 2014) |

Online type: $O(1/\sqrt{T})$ in general,
$O(\log(T)/T)$ or $O(1/T)$ for a strongly convex objective.
Batch type: linear convergence.

# Reminder of online stochastic optimization

$g_t \in \partial \ell_t(x_t),\ \bar{g}_t = \frac{1}{t}\sum_{\tau=1}^{t} g_\tau \quad (\ell_t(x) = \ell(z_t, x))$

**OPG (Online Proximal Gradient Descent)**

$$x_{t+1} = \arg\min_x \left\{ \langle g_t, x \rangle + \tilde{\psi}(x) + \frac{1}{2\eta_t}\|x - x_t\|^2 \right\}$$

**RDA (Regularized Dual Averaging; Xiao (2009), Nesterov (2009))**

$$x_{t+1} = \arg\min_x \left\{ \langle \bar{g}_t, x \rangle + \tilde{\psi}(x) + \frac{1}{2\eta_t}\|x\|^2 \right\}$$

These update rule is computed by a proximal mapping associated with $\tilde{\psi}$.

- Efficient for a simple regularization such as $\ell_1$ regularization.
- How about structured regularization? $\rightarrow$ ADMM.

# OPG-ADMM

Ordinary OPG: $x_{t+1} = \arg\min_x \left\{ \langle g_t, x \rangle + \tilde{\psi}(x) + \frac{1}{2\eta_t} \|x - x_t\|^2 \right\}.$

## OPG-ADMM

$$
\begin{aligned}
x_{t+1} =& \mathop{\mathrm{argmin}}_{x \in \mathcal{X}} \Big\{ {g_t}^\top x - {\lambda_t}^\top (B^\top x - y_t) \\
&\qquad + \frac{\rho}{2} \|B^\top x - y_t\|^2 + \frac{1}{2\eta_t} \|x - x_t\|_{G_t}^2 \Big\}, \\
y_{t+1} =& \mathop{\mathrm{argmin}}_{y \in \mathcal{Y}} \left\{ \psi(y) - \lambda_t^\top (B^\top x_{t+1} - y) + \frac{\rho}{2} \|B^\top x_{t+1} - y\|^2 \right\} \\
\lambda_{t+1} =& \lambda_t - \rho(B^\top x_{t+1} - y_{t+1}).
\end{aligned}
$$

- The update rule of $y_{t+1}$ and $\lambda_{t+1}$ are same as the ordinary ADMM.
- $\mathrm{prox}(\cdot|\psi)$ is usually analytically obtained.
- $G_t$ is any positive definite matrix.

# OPG-ADMM

Ordinary OPG: $x_{t+1} = \arg\min_x \left\{ \langle g_t, x \rangle + \tilde{\psi}(x) + \frac{1}{2\eta_t} \|x - x_t\|^2 \right\}.$

## OPG-ADMM

$$
\begin{aligned}
x_{t+1} =& \operatorname*{argmin}_{x \in \mathcal{X}} \Big\{ g_t^\top x - \lambda_t^\top (B^\top x - y_t) \\
& \quad + \frac{\rho}{2} \|B^\top x - y_t\|^2 + \frac{1}{2\eta_t} \|x - x_t\|_{G_t}^2 \Big\}, \\
y_{t+1} =& \operatorname*{argmin}_{y \in \mathcal{Y}} \left\{ \psi(y) - \lambda_t^\top (B^\top x_{t+1} - y) + \frac{\rho}{2} \|B^\top x_{t+1} - y\|^2 \right\} \\
=& \mathrm{prox}(B^\top x_{t+1} - \lambda_t/\rho | \psi), \\
\lambda_{t+1} =& \lambda_t - \rho(B^\top x_{t+1} - y_{t+1}).
\end{aligned}
$$

- The update rule of $y_{t+1}$ and $\lambda_{t+1}$ are same as the ordinary ADMM.
- $\mathrm{prox}(\cdot|\psi)$ is usually analytically obtained.
- $G_t$ is any positive definite matrix.

# RDA-ADMM

Ordinary RDA: $w_{t+1} = \arg\min_w \left\{ \langle \bar{g}_t, w \rangle + \tilde{\psi}(w) + \frac{1}{2\eta_t}\|w\|^2 \right\}$

## RDA-ADMM

Let $\bar{x}_t = \frac{1}{t}\sum_{\tau=1}^{t} x_\tau$, $\bar{\lambda}_t = \frac{1}{t}\sum_{\tau=1}^{t} \lambda_\tau$, $\bar{y}_t = \frac{1}{t}\sum_{\tau=1}^{t} y_\tau$, $\bar{g}_t = \frac{1}{t}\sum_{\tau=1}^{t} g_\tau$.

$$
\begin{aligned}
x_{t+1} =& \operatorname*{argmin}_{x\in\mathcal{X}} \Big\{ \bar{g}_t^\top x - (B\bar{\lambda}_t)^\top x + \frac{\rho}{2t}\|B^\top x\|^2 \\
& + \rho(B^\top \bar{x}_t - \bar{y}_t)^\top B^\top x + \frac{1}{2\eta_t}\|x\|_{G_t}^2 \Big\}, \\
y_{t+1} =& \mathrm{prox}(B^\top x_{t+1} - \lambda_t/\rho|\psi), \\
\lambda_{t+1} =& \lambda_t - \rho(B^\top x_{t+1} - y_{t+1}).
\end{aligned}
$$

The update rule of $y_{t+1}$ and $\lambda_{t+1}$ are same as the ordinary ADMM.

# Simplified version

Letting $G_t$ be a specific form, the update rule is much simplified.
($G_t = \gamma I - \frac{\rho}{t}\eta_t BB^\top$ for RDA-ADMM, and $G_t = \gamma I - \rho\eta_t BB^\top$ for OPG-ADMM)

## Online stochastic ADMM

1. Update of $x$

$$\text{(OPG-ADMM)} \quad x_{t+1} = \Pi_{\mathcal{X}}\left[-\frac{\eta_t}{\gamma}\{g_t - B(\lambda_t - \rho B^\top x_t + \rho y_t)\} + x_t\right],$$

$$\text{(RDA-ADMM)} \quad x_{t+1} = \Pi_{\mathcal{X}}\left[-\frac{\eta_t}{\gamma}\{\bar{g}_t - B(\bar{\lambda}_t - \rho B^\top \bar{x}_t + \rho \bar{y}_t)\}\right].$$

2. Update of $y$

$$y_{t+1} = \mathrm{prox}(B^\top x_{t+1} - \lambda_t/\rho|\psi).$$

3. Update of $\lambda$

$$\lambda_{t+1} = \lambda_t - \rho(B^\top x_{t+1} - y_{t+1}).$$

- Fast computation.
- Easy implementation.

# Convergence analysis

We bound the expected risk:

- Expected risk
$$P(x) = \mathrm{E}_Z[\ell(Z,x)] + \tilde{\psi}(x).$$

Assumptions:

(A1) $\exists G$ s.t. $\forall g \in \partial_x \ell(z,x)$ satisfies $\|g\| \leq G$ for all $z, x$.

(A2) $\exists L$ s.t. $\forall g \in \partial \psi(y)$ satisfies $\|g\| \leq L$ for all $y$.

(A3) $\exists R$ s.t. $\forall x \in \mathcal{X}$ satisfies $\|x\| \leq R$.

# Convergence rate: bounded gradient

(A1) $\exists G$ s.t. $\forall g \in \partial_x \ell(z, x)$ satisfies $\|g\| \le G$ for all $z, x$.
(A2) $\exists L$ s.t. $\forall g \in \partial \psi(y)$ satisfies $\|g\| \le L$ for all $y$.
(A3) $\exists R$ s.t. $\forall x \in \mathcal{X}$ satisfies $\|x\| \le R$.

**Theorem (Convergence rate of RDA-ADMM)**

*Under (A1), (A2), ~~(A3)~~, we have*

$$\mathrm{E}_{z_{1:T-1}}[P(\bar{x}_T) - P(x^*)] \le \frac{1}{T}\sum_{t=2}^{T} \frac{\eta_{t-1}}{2(t-1)} G^2 + \frac{\gamma}{\eta_T}\|x^*\|^2 + \frac{K}{T}.$$

**Theorem (Convergence rate of OPG-ADMM)**

*Under (A1), (A2), (A3), we have*

$$\mathrm{E}_{z_{1:T-1}}[P(\bar{x}_T) - P(x^*)] \le \tfrac{1}{2T}\textstyle\sum_{t=2}^{T} \max\left\{\tfrac{\gamma}{\eta_t} - \tfrac{\gamma}{\eta_{t-1}}, 0\right\} R^2 + \tfrac{1}{T}\textstyle\sum_{t=1}^{T} \tfrac{\eta_t}{2} G^2 + \tfrac{K}{T}.$$

Both methods have convergence rate $O\left(\frac{1}{\sqrt{T}}\right)$ by letting $\eta_t = \eta_0\sqrt{t}$ for RDA-ADMM and $\eta_t = \eta_0/\sqrt{t}$ for OPG-ADMM.

# Convergence rate: strongly convex

(A4) There exist $\sigma_f, \sigma_\psi \geq 0$ and $P, Q \succeq O$ such that

$$\mathrm{E}_Z[\ell(Z,x) + (x'-x)^\top \nabla_x \ell(Z,x)] + \frac{\sigma_f}{2}\|x - x'\|_P^2 \leq \mathrm{E}_Z[\ell(Z,x')],$$
$$\psi(y) + (y'-y)^\top \nabla \psi(y) + \frac{\sigma_\psi}{2}\|y - y'\|_Q^2,$$

for all $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$, and $\exists \sigma > 0$ satisfying

$$\sigma I \preceq \sigma_f P + \frac{\rho \sigma_\psi}{2\rho + \sigma_\psi} Q.$$

---

The update rule of RDA-ADMM is modified as

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \Big\{ \bar{g}_t^\top x - \bar{\lambda}_t^\top B^\top x + \frac{\rho}{2t}\|B^\top x\|^2 + \rho(B^\top \bar{x} t - \bar{y} t)^\top B^\top x + \frac{1}{2\eta_t}\|x\|_{G_t}^2$$
$$+ \frac{\sigma}{2}\|x - \bar{x}_t\|^2 \Big\}.$$

# Convergence analysis: strongly convex

**Theorem (Convergence rate of RDA-ADMM)**

*Under (A1), (A2), ~~(A3)~~, (A4), we have*

$$\mathrm{E}_{z_{1:T-1}}[P(\bar{x}_T) - P(x^*)] \leq \frac{1}{2T}\sum_{t=2}^{T}\frac{1}{\frac{t-1}{\eta_{t-1}} + t\sigma}G^2 + \frac{\gamma}{\eta_T}\|x^*\|^2 + \frac{K}{T}.$$

**Theorem (Convergence rate of OPG-ADMM)**

*Under (A1), (A2), (A3), (A4), we have*

$$\mathrm{E}_{z_{1:T-1}}[P(\bar{x}_T) - P(x^*)] \leq \frac{1}{2T}\sum_{t=2}^{T}\max\left\{\frac{\gamma}{\eta_t} - \frac{\gamma}{\eta_{t-1}} - \sigma, 0\right\}R^2 + \frac{1}{T}\sum_{t=1}^{T}\frac{\eta_t}{2}G^2 + \frac{K}{T}.$$

Both methods have convergence rate $O\left(\frac{\log(T)}{T}\right)$ by letting $\eta_t = \eta_0 t$ for RDA-ADMM and $\eta_t = \eta_0/t$ for OPG-ADMM.

This can be improved to $O(1/T)$ by weighted averaging (Azadi and Sra, 2014).

# Numerical experiments

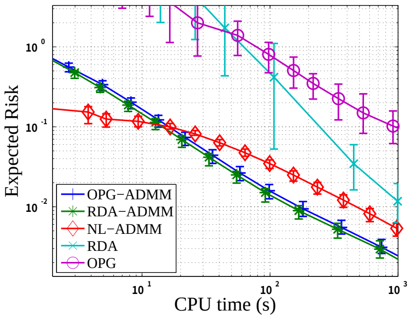



Figure: Artificial data: 1024 dim, 512 sample, Overlapped group lasso.

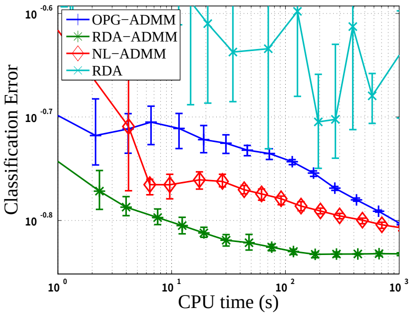



Figure: Real data (Adult, a9a): 15,252 dim, 32,561 sample, Overlapped group lasso + $\ell_1$ reg.

# Related methods

- $O(n/T)$ (improved from $O(1/\sqrt{T})$) convergence in a batch setting: Zhong and Kwok (2014)
- Acceleration of stochastic ADMM: Azadi and Sra (2014)
- Parallel computing with stochastic ADMM: Wang et al. (2014)

# Outline

1. Stochastic optimization for structured regularization
   - Structured regularization
   - Alternating Direction Method of Multipliers (ADMM)
   - Stochastic ADMM for online data
   - Stochastic ADMM for batch data
2. Parallel and distributed optimization
3. Further interesting topics

# Batch setting

In the batch setting, the data are fixed. We just minimize the objective function defined by

$$\frac{1}{n}\sum_{i=1}^{n} f_i(a_i^\top x) + \psi(B^\top x).$$

- ADMM version of SDCA
- Converges linearly:

$$T > (n + \gamma/\lambda)\log(1/\epsilon)$$

to achieve $\epsilon$ accuracy for $\gamma$-smooth loss and $\lambda$-strongly convex regularization.

# Dual problem

**Dual problem**

Let $A = [a_1, a_2, \dots, a_n] \in \mathbb{R}^{p \times n}$.

$$\min_w \left\{ \frac{1}{n} \sum_{i=1}^n f_i(a_i^\top w) + \psi(B^\top w) \right\} \qquad \text{(P: Primal)}$$

$$= - \min_{x \in \mathbb{R}^n, y \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^n f_i^*(x_i) + \psi^*\left(\frac{y}{n}\right) \;\middle|\; Ax + By = 0 \right\} \qquad \text{(D: Dual)}$$

Optimality condition:

$$a_i^\top w^* \in \nabla f_i^*(x_i^*), \;\; \frac{1}{n} y^* \in \nabla \psi(u)|_{u = B^\top w^*}, \;\; Ax^* + By^* = 0.$$

★ **Each coordinate $x_i$ corresponds to each observation $a_i$.**

# SDCA-ADMM

Let the augmented Lagrangian be
$\mathcal{L}(x, y, w) := \sum_{i=1}^{n} f_i^*(x_i) + n\psi^*(y/n) - \langle w, Ax + By\rangle + \frac{\rho}{2}\|Ax + By\|^2$.

**Basic algorithm**

For each $t = 1, 2, \ldots$
Choose $i \in \{1, \ldots, n\}$ uniformly at random, and update

$$y^{(t)} \leftarrow \arg\min_{y} \left\{ \mathcal{L}(x^{(t-1)}, y, w^{(t-1)}) + \frac{1}{2}\|y - y^{(t-1)}\|_Q^2 \right\}$$

$$x_i^{(t)} \leftarrow \arg\min_{x_i \in \mathbb{R}} \left\{ \mathcal{L}([x_i; x_{\setminus i}^{(t-1)}], y^{(t)}, w^{(t-1)}) + \frac{1}{2}\|x_i - x_i^{(t-1)}\|_{G_{i,i}}^2 \right\}$$

$$w^{(t)} \leftarrow w^{(t-1)} - \xi\rho\{n(Ax^{(t)} + By^{(t)}) - (n-1)(Ax^{(t-1)} + By^{(t-1)})\}.$$

$Q$, $G_{i,i}$ are positive definite matrices that satisfy some condition.

- Only $i$-th coordinate $x_i$ is updated.
- The update of the multiplier $w$ should be modified.

# SDCA-ADMM

Let the augmented Lagrangian be
$\mathcal{L}(x, y, w) := \sum_{i=1}^{n} f_i^*(x_i) + n\psi^*(y/n) - \langle w, Ax + By \rangle + \frac{\rho}{2}\|Ax + By\|^2$.

Split the index set $\{1, \dots, n\}$ into $K$ groups $(I_1, I_2, \dots, I_K)$.

**Block coordinate SDCA-ADMM**

For each $t = 1, 2, \dots$

Choose $k \in \{1, \dots, K\}$ uniformly at random, and set $I = I_k$,

$$y^{(t)} \leftarrow \arg\min_{y} \left\{ \mathcal{L}(x^{(t-1)}, y, w^{(t-1)}) + \frac{1}{2}\|y - y^{(t-1)}\|_Q^2 \right\}$$

$$x_I^{(t)} \leftarrow \arg\min_{x_I \in \mathbb{R}^{|I|}} \left\{ \mathcal{L}([x_I; x_{\setminus I}^{(t-1)}], y^{(t)}, w^{(t-1)}) + \frac{1}{2}\|x_I - x_I^{(t-1)}\|_{G_{I,I}}^2 \right\}$$

$$w^{(t)} \leftarrow w^{(t-1)} - \xi\rho\{n(Ax^{(t)} + By^{(t)}) - (n - n/K)(Ax^{(t-1)} + By^{(t-1)})\}.$$

$Q$, $G_{I,I}$ are positive definite matrices that satisfy some condition.

# Simplified algorithm

Setting
$$Q = \rho(\eta_B \mathrm{I}_d - B^\top B),\ \ G_{I,I} = \rho(\eta_{Z,I}\mathrm{I}_{|I|} - Z_I^\top Z_I),$$
then, by the relation $\mathrm{prox}(q|\psi) + \mathrm{prox}(q|\psi^*) = q$, we obtain the following update rule:

**Simplified algorithm**

- For $q^{(t)} = y^{(t-1)} + \frac{B^\top}{\rho\eta_B}\{w^{(t-1)} - \rho(Zx^{(t-1)} + By^{(t-1)})\}$, let
$$y^{(t)} \leftarrow q^{(t)} - \mathrm{prox}(q^{(t)}|n\psi(\rho\eta_B\ \cdot\ )/(\rho\eta_B)),$$
- For $p_I^{(t)} = x_I^{(t-1)} + \frac{Z_I^\top}{\rho\eta_{Z,I}}\{w^{(t-1)} - \rho(Zx^{(t-1)} + By^{(t)})\}$, let
$$x_i^{(t)} \leftarrow \mathrm{prox}(p_i^{(t)}|f_i^*/(\rho\eta_{Z,I})) \qquad (\forall i \in I).$$

★ The update of $x$ can be parallelized.

# Convergence Analysis

$x^*$: the optimal variable of $x$
$\mathcal{Y}^*$: the set of optimal variables (not necessarily unique)
$w^*$: the optimal variable of $w$
**Assumption**:

- There exits $v > 0$ such that, $\forall x_i \in \mathbb{R}$,

$$f_i^*(x_i) - f_i^*(x_i^*) \geq \langle \nabla f_i^*(x_i^*), x_i - x_i^* \rangle + \frac{\|x_i - x_i^*\|^2}{2\gamma}.$$

  $\exists h, v_\psi > 0$ such that, for all $y, u$, there exists $\widehat{y}^* \in \mathcal{Y}^*$ such that

$$\psi^*(y/n) - \psi^*(\widehat{y}^*/n) \geq \langle B^\top w^*, y/n - \widehat{y}^*/n \rangle + \frac{v_\psi}{2} \|P_{\mathrm{Ker}(B)}(y/n - \widehat{y}^*/n)\|^2,$$

$$\psi(u) - \psi(B^\top w^*) \geq \langle y^*/n, u - B^\top w^* \rangle + \frac{\lambda}{2} \|u - B^\top w^*\|^2.$$

- $B^\top$ is injective.

$F_D(x,y) := \frac{1}{n}\sum_{i=1}^n f_i^*(x_i) + \psi^*(\frac{y}{n}) - \langle w^*, A\frac{x}{n} - B\frac{y}{n}\rangle$.
$R_D(x,y,w) := F_D(x,y) - F_D(x^*,y^*) + \frac{1}{2n^2\xi\rho}\|w-w^*\|^2 + \frac{\rho(1-\xi)}{2n}\|Ax + By\|^2 + \frac{1}{2n}\|x-x^*\|^2_{\mathrm{I}_p/\gamma+H} + \frac{1}{2nK}\|y-y^*\|^2_Q$.

**Theorem (Linear convergence of SDCA-ADMM)**

*Let $H$ be a matrix such that $H_{I,I} = \rho A_I^\top A_I + G_{I,I}$ for all $I \in \{I_1,\dots,I_K\}$,*

$$\mu = \min\left\{\frac{1}{4(1+\gamma\sigma_{\max}(H))}, \frac{\lambda\rho\sigma_{\min}(B^\top B)}{2\max\{1/n, 4\lambda\rho, 4\lambda\sigma_{\max(Q)}\}}, \frac{K v_\psi/n}{4\sigma_{\max}(Q)}, \frac{K\sigma_{\min}(BB^\top)}{4\sigma_{\max(Q)}(\rho\gamma\sigma_{\max}(A^\top A)+4)}\right\},$$

$\xi = \frac{1}{4n}$, *and* $C_1 = R_D(x^{(0)}, y^{(0)}, w^{(0)})$, *then we have that*

$$\text{(dual residual)}\quad \mathrm{E}[R_D(x^{(t)}, y^{(t)}, w^{(t)})] \le \left(1 - \frac{\mu}{K}\right)^t C_1,$$

$$\text{(primal variable)}\quad \mathrm{E}[\|w^{(t)} - w^*\|^2] \le \frac{n\rho}{2}\left(1 - \frac{\mu}{K}\right)^t C_1.$$

For $t \ge C'\frac{K}{\mu}\log\left(\frac{C''n}{\epsilon}\right)$, we have that $\mathrm{E}[\|w^{(t)} - w^*\|^2] \le \epsilon$.

# Convergence Analysis

**Assumption:**

- $f_i$ is $\gamma$-smooth.
- $\tilde{\psi}$ is $\lambda$-strongly convex.
- Other technical conditions.

With a setting $\rho = \min\{1, 1/\gamma\}$ and $K = n$,

$$t \geq C\left(n + \frac{\gamma}{\lambda}\right) \log\left(\frac{C'}{\epsilon}\right)$$

gives $\mathrm{E}[\|w^{(t)} - w^*\|^2] \leq \epsilon$ ($\frac{\gamma}{\lambda}$ is like the condition number).
The rate is as good as the ordinary SDCA.

---

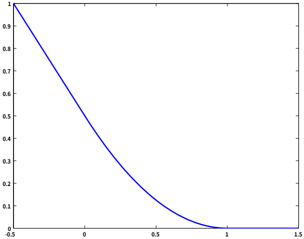
Non stochastic method (e.g. ADMM in dual):

$$t \geq C n \frac{\gamma}{\lambda} \log\left(\frac{C'}{\epsilon}\right).$$

# Numerical Experiments: Loss function

Binary classification.

Smoothed hinge loss: $f_i(u) = \begin{cases} 0, & (y_i u \geq 1), \\ \frac{1}{2} - y_i u, & (y_i u < 0), \\ \frac{1}{2}(1 - y_i u)^2, & (\text{otherwise}). \end{cases}$




⇒ proximal mapping is analytically obtained.

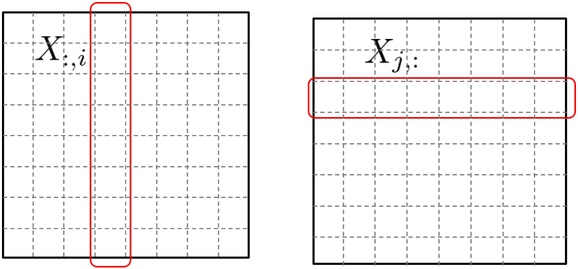
$$\mathrm{prox}(u|f_i^*/C) = \begin{cases} \frac{Cu - y_i}{1+C} & (-1 \leq \frac{Cuy_i - 1}{1+C} \leq 0), \\ -y_i & (-1 > \frac{Cuy_i - 1}{1+C}), \\ 0 & (\text{otherwise}). \end{cases}$$

# Numerical Experiments (Artificial data): Setting

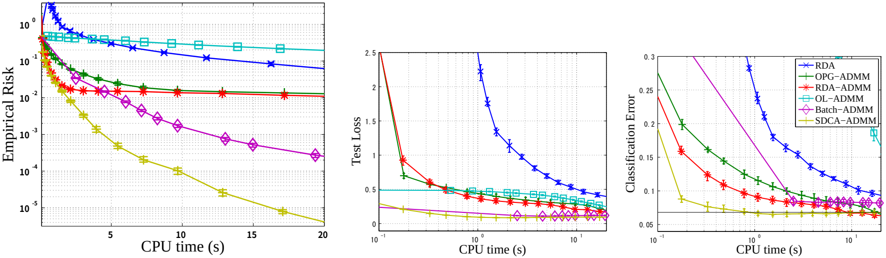
**Artificial data**: Overlapped group regularization:

$$\tilde{\psi}(X) = C(\sum_{i=1}^{32} \|X_{:,i}\| + \sum_{j=1}^{32} \|X_{j,:}\| + 0.01 \times \sum_{i,j} X_{i,j}^2/2),$$

$X \in \mathbb{R}^{32 \times 32}$.



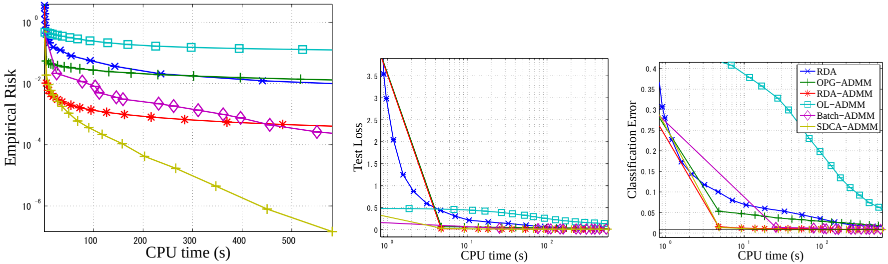
# Numerical Experiments (Artificial data): Results




(a) Excess objective value (b) Test loss (c) Class. Error

Figure: Artificial data ($n$=5,120, $d$=1024). Overlapped group lasso. Mini-batch size $n/K = 50$.

# Numerical Experiments (Artificial data): Results




(a) Excess objective value (b) Test loss (c) Class. Error

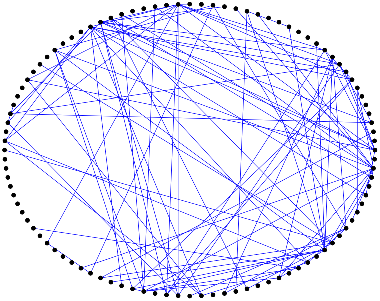
Figure: Artificial data ($n$=51,200, $d$=1024). Overlapped group lasso. Mini-batch size $n/K = 50$.

# Numerical Experiments (Real data): Setting

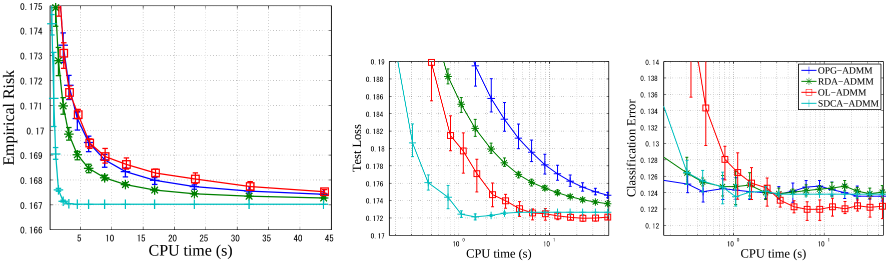
Real data: Graph regularization:

$$\tilde{\psi}(w) = C_1 \sum_{i=1}^{p} |w_i| + C_2 \sum_{(i,j)\in E} |w_i - w_j| + 0.01 \times (C_1 \sum_{i=1}^{p} |w_i|^2 + C_2 \sum_{(i,j)\in E} |w_i - w_j|^2)$$

where $E$ is the set of edges obtained from the similarity matrix.
The similarity graph is obtained by Graphical Lasso (Yuan and Lin, 2007, Banerjee et al., 2008).



# Numerical Experiments (Real data): Results



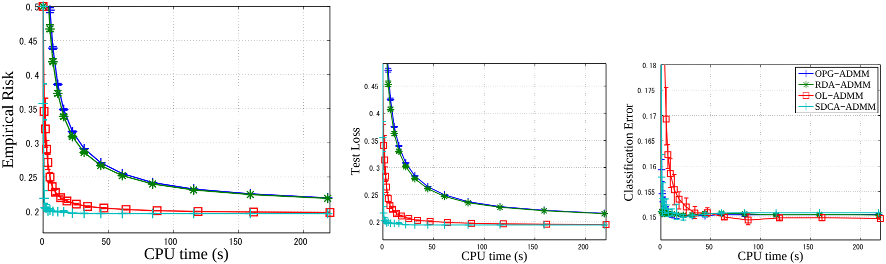

(a) Objective function

(b) Test loss

(c) Class. Error

Figure: Real data (20 Newsgroups, $n$=12,995). Graph regularization. Mini-batch size is $n/K = 50$.

# Numerical Experiments (Real data): Results




(a) Objective function (b) Test loss (c) Class. Error

Figure: Real data (a9a, $n$=32,561). Graph regularization. Mini-batch size $n/K = 50$.

# Outline


1. Stochastic optimization for structured regularization
   - Structured regularization
   - Alternating Direction Method of Multipliers (ADMM)
   - Stochastic ADMM for online data
   - Stochastic ADMM for batch data
2. Parallel and distributed optimization
3. Further interesting topics

# Distributed computing

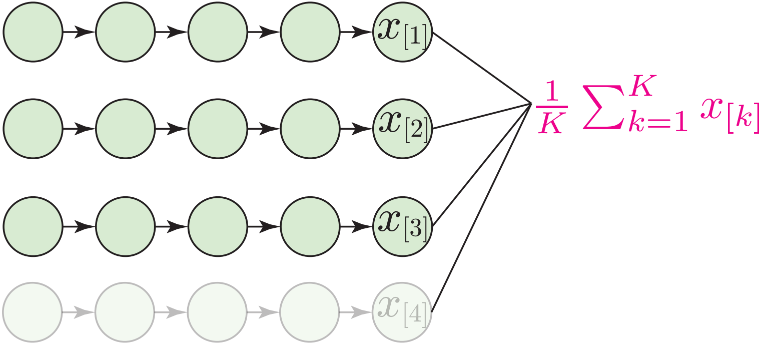
Q: How to deal with huge data that cannot be treated in one computational node?
A: Distributed computing.

**Challenge** communication cost trade-off:
communication inside node v.s. between nodes via network.

We briefly introduce two approaches

- Simple averaging of SGD: (Zinkevich et al., 2010, Zhang et al., 2013)
- Distributed dual coordinate descent (COCOA+): (Ma et al., 2015)

# Simple averaging




- Run independent SGDs by $K$ nodes.
- Take the average of $K$ final solutions:

$$\hat{x}_K = \frac{1}{K}\sum_{k=1}^{K} x_{[k]}.$$

Just one synchronization, efficient communication cost. How about convergence?

**Assumption:**

- The loss function is sufficiently smooth (there exists second order derivative, and it is Lipschitz continuous and bounded).
- The expected loss is $\lambda$-**strongly convex**.
- **Each node runs $T$ iterations of SGD**.

**Theorem ((Zhang et al., 2013))**

*With an appropriate step size,*

$$\mathrm{E}[\|\hat{x}_K - x^*\|^2] \leq C\left(\frac{G^2}{KT\lambda^2} + \frac{1}{T^{3/2}}\right).$$

- As $K$ is increased, the main term is linearly improved.
- However, too large $K$ is not effective. Actually, for $\lambda \in (0, 1/\sqrt{T})$, it is shown that (Shamir et al., 2014)

$$\mathrm{E}[\|\hat{x}_K - x^*\|^2] \geq \frac{C}{\lambda^2 T}.$$

# COCOA+ (Ma et al., 2015)

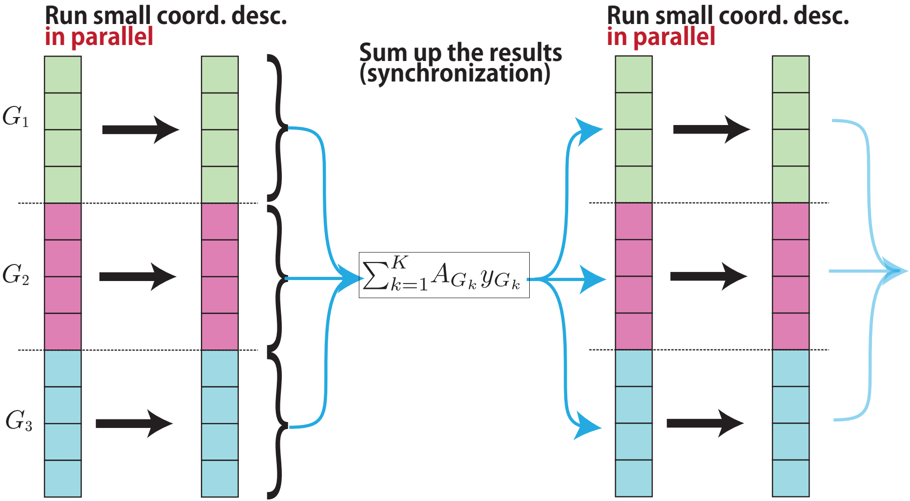
We divide the sample into $K$ groups $\{G_k\}_k$:

$$\{1, \ldots, n\} = \bigcup_{k=1}^{K} G_k, \;\; G_k \cap G_{k'} = \emptyset.$$

**Dual problem of RERM**:

$$\begin{aligned} D(y) =& \frac{1}{n} \sum_{i=1}^{n} f_i^*(y_i) + \psi^* \left( -\frac{1}{n} A y \right) \\ =& \frac{1}{n} \sum_{k=1}^{K} \underbrace{\left( \sum_{i \in G_k} f_i^*(y_i) \right)}_{\text{Divided into } K \text{ groups}} + \underbrace{\psi^* \left( -\frac{1}{n} \sum_{k=1}^{K} A_{G_k} y_{G_k} \right)}_{\text{needs synchronization}} \end{aligned}$$

where $A_{G_k} = [a_{i_1}, \ldots, a_{i_{G_k}}] \in \mathbb{R}^{p \times |G_k|}$ where $i_j \in G_k$ and $y_{G_k} = (y_i)_{i \in G_k}$.


Run small coord. desc.
in parallel
Sum up the results
(synchronization)
Run small coord. desc.
in parallel
$G_1$
$G_2$
$G_3$
$\sum_{k=1}^{K} A_{G_k} y_{G_k}$

1. $f_i$ is $\gamma_f$-smooth.
2. $\psi$ is $\lambda$-strongly convex.
3. Each subproblem decreases the objective by a factor of $\Theta$.

Let the separability of the problems as

$$\sigma_{\min} := \max_{y \in \mathbb{R}^n} \frac{\|Ay\|^2}{\sum_{k=1}^{K} \|A_{G_k} y_{G_k}\|^2}, \quad \sigma_{\max} := \max_{k} \max_{y_{G_k} \in \mathbb{R}^{|G_k|}} \frac{\|A_{G_k} y_{G_k}\|^2}{\|y_{G_k}\|^2}.$$

## Theorem

*Under an appropriate setting of parameters, after $T$ iterations with*

$$T \geq C \frac{\sigma_{\min}\sigma_{\max}\gamma_f/(\lambda n)+1}{(1-\Theta)} \log(1/\epsilon),$$

*it holds that* $\mathrm{E}[D(y^{(T)}) - D(y^*)] \leq \epsilon$. *Furthermore, after $T$ iterations with*

$$T \geq C \frac{\sigma_{\min}\sigma_{\max}\gamma_f/\lambda+1}{(1-\Theta)} \log\left(\frac{\sigma_{\min}\sigma_{\max}\gamma_f/\lambda+1}{(1-\Theta)}/\epsilon\right),$$

*it hods that* $\mathrm{E}[P(w^{(t)}) - D(y^{(t)})] \leq \epsilon$.

It is shown that $\sigma_{\text{min}} \leq K$ and $\sigma_{\text{max}} \leq n/K$. Then

$$T \geq C \frac{\gamma_f/\lambda + 1}{(1 - \Theta)} \log(1/\epsilon)$$

achieves $\epsilon$ accuracy. This is equivalent to iteration number of batch gradient methods on a strongly convex function.

Typically $\Theta = \left(1 - \frac{1}{n/K + \gamma_f/\lambda}\right)^t$ where $t$ is the number of inner iterations.
Total computational time (worst case):

$$t(1 + \gamma_f/\lambda) \log(1/\epsilon) = \left(\frac{n}{K} + \frac{\gamma_f}{\lambda}\right) \left(1 + \frac{\gamma_f}{\lambda}\right) \log(1/\epsilon).$$

Huge learning problems can be optimized on a distributed system with **linear convergence rate**.

# Outline


1. Stochastic optimization for structured regularization
    - Structured regularization
    - Alternating Direction Method of Multipliers (ADMM)
    - Stochastic ADMM for online data
    - Stochastic ADMM for batch data
2. Parallel and distributed optimization
3. Further interesting topics

# Stochastic primal dual coordinate method

**(Zhang and Lin, 2015)**

$$\min_x \left\{ \frac{1}{n} \sum_{i=1}^{n} \underbrace{f_i(a_i^\top x)} + \psi(x) \right\}$$

$x$ **and** $y_i$ **(**$i$ **is chosen randomly) are updated alternatively.**
Iteration complexity:

$$T \geq \left( n + \sqrt{\gamma/\lambda} \right) \log(1/\epsilon)$$

# Stochastic primal dual coordinate method

### (Zhang and Lin, 2015)

$$\min_{x}\left\{\frac{1}{n}\sum_{i=1}^{n}\underbrace{f_i(a_i^\top x)}_{\sup_{y_i}\{\langle x, a_i y_i\rangle - f_i^*(y_i)\}} + \psi(x)\right\}$$

$$= \min_{x}\max_{y}\left\{\frac{1}{n}\sum_{i=1}^{n}(\langle x, a_i y_i\rangle - f_i^*(y_i)) + \psi(x)\right\}$$

**$x$ and $y_i$ ($i$ is chosen randomly) are updated alternatively.**
Iteration complexity:

$$T \geq \left(n + \sqrt{\gamma/\lambda}\right)\log(1/\epsilon)$$

# Multi-armed bandit


Maximize the sum of rewards earned through a sequence of plays.

- Formulated by Robbins (1952).
- Optimal strategy: Lai and Robbins (1985)
- UCB strategy: Auer et al. (2002), Burnetas and Katehakis (1996)
- Thompson sampling (Bayesian strategy): Thompson (1933)

Continuous version of bandit: **Bayesian optimization** (Močkus, 1975, Mockus and Mockus, 1991, Srinivas et al., 2012, Snoek et al., 2012)

- Gaussian process regression to search the peak of a function.
- Practically useful for hyper-parameter tuning of deep learning.

# Stochastic gradient Langevin Monte Carlo

## (Welling and Teh, 2011)

**Goal:** Efficient sampling from the posterior distribution.

Randomly choose small mini-batch $I_t \subseteq \{1, \dots, n\}$ and update the parameter $\theta_t$:

$$\theta_t = \theta_{t-1} + \eta_t \left[ \nabla_\theta \underbrace{\log \pi(\theta)}_{\text{prior}} + \frac{n}{|I_t|} \sum_{i \in I_t} \nabla_\theta \underbrace{\log(p(x_i|\theta))}_{\text{likelihood}} \right] + \epsilon_t \xi$$

where $\xi \sim N(0, \eta_t \mathrm{I})$ and

$$\textstyle\sum_{t=1}^{\infty} \epsilon_t = \infty, \qquad \sum_{t=1}^{\infty} \epsilon_t^2 < \infty.$$

Related topic: Stochastic variational inference (Hoffman et al., 2013).

- Bayesian inference for large datasets.
- Practically very useful.
- Supported by some theoretical backgrounds.

# Connection to learning theory

$$\min_x \left\{ \frac{1}{t} \sum_{\tau=1}^{t} \ell(z_\tau, x) + \lambda_t \|x\|_1 \right\}$$

- Solve regularized learning problem in online setting (data $z_t$ comes one after another).
- The regularization parameter $\lambda_t$ should go to zero as $t \to \infty$.

**Question: Can we achieve statistically optimal estimation?**
$\to$ **Yes**.

- RADAR ($L_1$-regularization): Agarwal et al. (2012)
- REASON (Stochastic ADMM): Sedghi et al. (2014)

$$\|x_t - x^*\|^2 \leq \frac{s \log(p)}{T}$$

for $s$-sparse truth $x^*$.
**Simultaneous discussions of optimization and statistics.**

# Summary of part III

- Stochastic ADMM for structured sparsity
  - Online proximal gradient method, regularized dual averaging method for ADMM (online)
  - Stochastic dual coordinate descent for ADMM (batch)
  - A similar convergence result to the normal ones
- Distributed stochastic optimization
  - Simple averaging
  - Distributed SDCA (COCOA+)

A. Agarwal, S. Negahban, and M. J. Wainwright. Stochastic optimization and sparse statistical recovery: Optimal algorithms for high dimensions. In Advances in Neural Information Processing Systems, pages 1538–1546, 2012.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. Machine learning, 47(2-3):235–256, 2002.

S. Azadi and S. Sra. Towards an optimal stochastic alternating direction method of multipliers. In Proceedings of the 31st International Conference on Machine Learning, pages 620–628, 2014.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. Foundations and Trends® in Machine Learning, 4(1):1–106, 2012.

F. R. Bach. Structured sparsity-inducing norms through submodular functions. In Advances in Neural Information Processing Systems, pages 118–126, 2010.

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.

A. N. Burnetas and M. N. Katehakis. Optimal adaptive policies for sequential allocation problems. Advances in Applied Mathematics, 17 (2):122–142, 1996.

A. Chambolle. An algorithm for total variation minimization and applications. Journal of Mathematical imaging and vision, 20(1-2): 89–97, 2004.

W. Deng and W. Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. Technical report, Rice University CAAM TR12-14, 2012.

J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. Journal of Machine Learning Research, 10: 2873–2908, 2009.

D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite-element approximations. Computers & Mathematics with Applications, 2:17–40, 1976.

B. He and X. Yuan. On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. SIAM J. Numerical Analisis, 50(2):700–709, 2012.

M. Hestenes. Multiplier and gradient methods. Journal of Optimization Theory & Applications, 4:303–320, 1969.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. The Journal of Machine Learning Research, 14(1): 1303–1347, 2013.

M. Hong and Z.-Q. Luo. On the linear convergence of the alternating direction method of multipliers. arXiv preprint arXiv:1208.3922, 2012.

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`.

Y. Kawahara, K. Nagano, K. Tsuda, and J. A. Bilmes. Submodularity cuts and applications. In Advances in Neural Information Processing Systems, pages 916–924, 2009.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. Advances in applied mathematics, 6(1):4–22, 1985.

N. Le Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite training sets. In Advances in Neural Information Processing Systems 25, 2013.

C. Ma, V. Smith, M. Jaggi, M. Jordan, P. Richtárik, and M. Takáč. Adding vs. averaging in distributed primal-dual optimization. In Proceedings of The 32nd International Conference on Machine Learning (ICML-15), pages 1973–1982, 2015.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Non-local sparse models for image restoration. In Computer Vision, 2009 IEEE 12th International Conference on, pages 2272–2279. IEEE, 2009.

J. Močkus. On bayesian methods for seeking the extremum. In Optimization Techniques IFIP Technical Conference, pages 400–404. Springer, 1975.

J. Mockus and L. Mockus. Bayesian approach to global optimization and application to multiobjective and constrained problems. Journal of Optimization Theory and Applications, 70(1):157–172, 1991.

J. F. Mota, J. M. Xavier, P. M. Aguiar, and M. Püschel. A proof of convergence for the alternating direction method of multipliers applied to polyhedral-constrained functions. arXiv preprint arXiv:1112.2295, 2011.

Y. Nesterov. Gradient methods for minimizing composite objective function. Technical Report 76, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), 2007.

Y. Nesterov. Primal-dual subgradient methods for convex problems. Mathematical Programming, Series B, 120:221–259, 2009.

H. Ouyang, N. He, L. Q. Tran, and A. Gray. Stochastic alternating direction method of multipliers. In Proceedings of the 30th International Conference on Machine Learning, 2013.

M. Powell. A method for nonlinear constraints in minimization problems. In R. Fletcher, editor, Optimization, pages 283–298. Academic Press, London, New York, 1969.

H. Robbins. Some aspects of the sequential design of experiments. Bulletin of the American Mathematical Society, 58:527–535, 1952.

R. T. Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. Mathematics of Operations Research, 1:97–116, 1976.

L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena, 60(1):259–268, 1992.

H. Sedghi, A. Anandkumar, and E. Jonckheere. Multi-step stochastic admm in high dimensions: Applications to sparse optimization and matrix decomposition. In Advances in Neural Information Processing Systems, pages 2771–2779, 2014.

S. Shalev-Shwartz and T. Zhang. Proximal stochastic dual coordinate ascent. Technical report, 2013. arXiv:1211.2717.

O. Shamir, N. Srebro, and T. Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In Proceedings of the 32th International Conference on Machine Learning, pages 1000–1008, 2014.

J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In Advances in neural information processing systems, pages 2951–2959, 2012.

N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. Information Theory, IEEE Transactions on, 58(5): 3250–3265, 2012.

T. Suzuki. Dual averaging and proximal gradient descent for online alternating direction multiplier method. In Proceedings of the 30th International Conference on Machine Learning, pages 392–400, 2013.

T. Suzuki. Stochastic dual coordinate ascent with alternating direction method of multipliers. In Proceedings of the 31th International Conference on Machine Learning, pages 736–744, 2014.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, pages 285–294, 1933.

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.

R. J. Tibshirani and J. Taylor. The solution path of the generalized lasso. Ann. Statist., 39(3):1335–1371, 06 2011. doi: 10.1214/11-AOS878. URL http://dx.doi.org/10.1214/11-AOS878.

H. Wang and A. Banerjee. Online alternating direction method. In Proceedings of the 29th International Conference on Machine Learning, 2012.

H. Wang, A. Banerjee, and Z.-Q. Luo. Parallel direction method of multipliers. In Advances in Neural Information Processing Systems, pages 181–189, 2014.

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient langevin dynamics. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), pages 681–688, 2011.

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.

Y.-L. Yu. On decomposing the proximal map. In Advances in Neural Information Processing Systems, pages 91–99, 2013.

L. Yuan, J. Liu, and J. Ye. Efficient methods for overlapping group lasso. In Advances in Neural Information Processing Systems 24, 2011.

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

Y. Zhang and X. Lin. Stochastic primal-dual coordinate method for regularized empirical risk minimization. In Proceedings of The 32nd International Conference on Machine Learning (ICML-15), pages 353–361, 2015.

Y. Zhang, J. C. Duchi, and M. Wainwright. Communication-efficient algorithms for statistical optimization. Journal of Machine Learning Research, 14:3321–3363, 2013.

W. Zhong and J. Kwok. Fast stochastic alternating direction method of multipliers. In Proceedings of the 31st International Conference on Machine Learning (ICML-14), pages 46–54, 2014.

M. Zinkevich, M. Weimer, L. Li, and A. J. Smola. Parallelized stochastic gradient descent. In Advances in neural information processing systems, pages 2595–2603, 2010.