3. Using Sparse Coding for Multitask Learning and Learning to Learn

Massimiliano Pontil

May 2015
Plan

- Problem setting and method
- Bound for learning to learn
- Bound for MTL
- Bounding the total $\ell_1$ norm
- Subspace learning

Main reference: Maurer, Pontil, Romera-Paredes. Sparse coding for multitask and transfer learning. ICML 2013
Learning Sparse Representations

- Represent the regression vectors as combinations of some vectors (called dictionary elements or atoms):

\[ w_t = D \gamma_t = \sum_{k=1}^{K} D_k \gamma_{kt} \]

- Set of dictionaries \( \mathcal{D}_K := \left\{ D \in \mathbb{R}^{d \times K} : \max_{k=1}^{K} \| D_k \|_2 \leq 1 \right\} \)

- Learning method:

\[
\min_{D \in \mathcal{D}_K} \frac{1}{T} \sum_{t=1}^{T} \min_{\gamma \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} \ell \left( \langle D \gamma, x_{ti} \rangle, y_{ti} \right)
\]

- \( \mathcal{C} \) is a bounded subset of \( \mathbb{R}^d \)
Choose $C = \{ \| \gamma \|_1 \leq \alpha \}$, so to encourage $w_t$ to be **sparse combinations** of the dictionary elements.

For fixed $D$ each regression vector is estimated by the Lasso with **feature map** $\phi(x) = D^* x$, that is we run the Lasso on the preprocessed data $(D^* x_1, y_1), ..., (D^* x_n, y_n)$.

Two regularization parameters: $K$ and $\alpha$.

Intuition: weaker way to relate the tasks: a pair of tasks $s$ and $t$ will typically share only few dictionary elements since $\gamma_t$ and $\gamma_s$ tend to be sparse.
Training:

- for $t = 1, \ldots, T$ draw $\mu_t \sim \mathcal{E}$ then draw a dataset $z_t \sim (\mu_t)^n$
- learn a dictionary $\hat{D}$ by solving

$$\min_{D \in \mathcal{D}_K} \frac{1}{T} \sum_{t=1}^{T} \min_{\gamma \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} \ell(\langle D\gamma, x_{ti} \rangle, y_{ti})$$

Testing:

- draw a task $\mu \sim \mathcal{E}$, then draw a dataset $z \sim \mu^n$
- run the Lasso with dictionary $\hat{D}$:

$$\min_{\gamma \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} \ell(\langle \hat{D}\gamma, x_i \rangle, y_i)$$
LTL - recap II

The notation $\mathbf{z} \sim \hat{\mathcal{E}}$ means the draw of a random $n$-dataset from the environment: we first sample a task $\mu \sim \mathcal{E}$ then sample

$$\mathbf{z} := (x_1, y_1), \ldots, (x_n, y_n) \sim \mu^n$$

Hence

$$\mathbb{E}_{\mathbf{z} \sim \hat{\mathcal{E}}} [\cdot] := \mathbb{E}_{\mu \sim \mathcal{E}} \mathbb{E}_{\mathbf{z} \sim \mu^n} [\cdot]$$

The error of a dictionary $D$ on a random dataset is

$$R_n(D) = \mathbb{E}_{\mathbf{z} \sim \hat{\mathcal{E}}} \mathbb{E}_{(x,y) \sim \mu} \ell(\langle D \gamma_D(z), x \rangle, y)$$

where

$$\gamma_D(z) = \arg\min_{\gamma \in \mathcal{C}} \sum_{i=1}^n \ell(\langle D \gamma, x_i \rangle, y_i)$$
LTL - recap III

For an infinite sample the error of a dictionary becomes

\[ R(D) = \mathbb{E} \min_{\mu \sim \mathcal{E}} \mathbb{E}_{\gamma \in \mathcal{C}} \mathbb{E}_{(x, y) \sim \mu} \ell(\langle D\gamma, x \rangle, y) \]

The error of the best dictionary given complete knowledge of the environment is then

\[ R^* := \min_{D \in \mathcal{D}_K} R(D) \]

Our goal is to compare the performance of the dictionary \( \hat{D} \) learned from the training datasets to the error of the best dictionary

\[ R_n(\hat{D}) - R^* \]

We call this the transfer error
Square loss

Consider the environment where the draw of $\mu$ is identified with the draw of a noiseless linear regression model: $\mu(x, y) = \delta(y - \langle w, x \rangle)p(x)$. If $p$ is the standard Gaussian in $d$ dimensions

$$\mathbb{E}_{(x, y) \sim \mu} \ell(\langle D \gamma, x \rangle, y) = \mathbb{E}_{x \sim p}(\langle D \gamma - w, x \rangle)^2 = \|w - D \gamma\|^2$$

so we recover the risk for dictionary learning (lecture 1)

$$R(D) = \mathbb{E}_{w \sim \mathcal{E}} \min_{\gamma} \|w - D \gamma\|^2$$

we also have

$$\hat{R}_D(z) = \min_{\gamma} \frac{1}{n} \sum_{i=1}^{n} \ell(\langle D \gamma, x_i \rangle, y_i) = \min_{\gamma} \|w - D \gamma\|^2 \hat{c}(x)$$

and denoting by $\gamma_D(w, x)$ a minimizer we obtain

$$R_n(D) = \mathbb{E}_{w \sim \mathcal{E}} \mathbb{E}_{x \sim p^n} \|w - D \gamma_D(w, x)\|^2$$
We recall some notation and introduced more we will use below

- $H$: Hilbert space (in some cases $\mathbb{R}^d$)
- input sequence $\mathbf{x} = (x_1, \ldots, x_n) \in H^n$
- $\mathbf{z} = (((x_1, y_1), \ldots, (x_n, y_n))$: sample or training set
- $\mathbf{z}_t = (((x_{t1}, y_{t1}), \ldots, (x_{tn}, y_{tn}))$: sample for the $t$-th task
- $\mathbf{x}_t = (x_{t1}, \ldots, x_{tn})$
- $\mathbf{X} = (x_1, \ldots, x_T)$
Covariance matrices

- $\hat{\Sigma}(x) = \frac{1}{m} \sum_{i=1}^{n} x_i x_i^T$: the empirical covariance
- $\hat{\Sigma}_t = \Sigma(x_t)$: the empirical covariance for $t$-th dataset
- $\hat{C} = \frac{1}{Tn} \sum_{t} x_{ti} x_{ti}^T$: total empirical covariance
- If $M$ is a matrix $\|M\|_p$ for $p \in [1, \infty]$ is the $p$-Schatten norm ($\ell_p$ norm of the singular values)
- $\hat{S}_p \equiv \hat{S}_p(X) = \frac{1}{T} \sum_{t} \|\hat{\Sigma}_t\|_p$
**Theorem 1.** Let $S_{\infty}(\mathcal{E}) := \mathbb{E}_{\mu \sim \mathcal{E}} \mathbb{E}_{(x,y) \sim \mu^n} \|\Sigma(x)\|_{\infty}$. With pr. $\geq 1 - \delta$

$$R_n(\hat{D}) - R^* \leq 4L\alpha \sqrt{\frac{S_{\infty}(\mathcal{E})(2 + \ln K)}{n}} + L\alpha K \sqrt{\frac{2\pi \hat{S}_1}{T}} + \sqrt{\frac{8 \ln \frac{4}{\delta}}{T}}$$

The key step in the bound is to bound the uniform deviation

$$\sup_{D \in \mathcal{D}} \left\{ R(D) - \frac{1}{T} \sum_{t=1}^{T} \min_{\gamma \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} \ell(\langle D_{\gamma}, x_{ti} \rangle, y_{ti}) \right\}$$

We will use the shorthand $\hat{R}_D(z_t) = \min_{\gamma \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} \ell(\langle D_{\gamma}, x_{ti} \rangle, y_{ti})$
Bound for learning to learn

Let $D_*$ and $\gamma_\mu$ be the minimizers in the definition of $R^*$, so that

$$R^* = \mathbb{E}_{\mu \sim \mathcal{E}} \mathbb{E}_{(x,y) \sim \mu} \ell(\langle D_* \gamma_\mu, x \rangle, y)$$

Decompose the transfer error as the sum of four terms

$$R_n(\hat{D}) - R^* = R_n(\hat{D}) - \frac{1}{T} \sum_{t=1}^{T} \hat{R}_{D}(z_t)$$

$$+ \frac{1}{T} \sum_{t=1}^{T} \hat{R}_{D}(z_t) - \frac{1}{T} \sum_{t=1}^{T} \hat{R}_{D_*}(z_t)$$

$$+ \frac{1}{T} \sum_{t=1}^{T} \hat{R}_{D_*}(z_t) - \mathbb{E}_{z \sim \hat{E}} \hat{R}_{D_*}(z)$$

$$+ \mathbb{E}_{z \sim \hat{E}} \hat{R}_{D_*}(z) - R^*$$
Bound for Learning to Learn

- The term (2) is non-positive by the definition of $\hat{D}$
- The term (4) is also non-positive because

$$\mathbb{E}_{z \sim \mu^n} \hat{R}_{D_*}(z) = \mathbb{E}_{z \sim \mu^n} \min_{\gamma \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} \ell(\langle D_* \gamma, x_i \rangle, y_i)$$

$$\leq \mathbb{E}_{z \sim \mu^n} \frac{1}{n} \sum_{i=1}^{n} \ell(\langle D_* \gamma, x_i \rangle, y_i) = \mathbb{E}_{(x,y) \sim \mu} \ell(\langle D_* \gamma, x \rangle, y)$$

- By Hoeffding's inequality the term (3) is less than $\sqrt{\ln(2/\delta)/2T}$ with probability at least $1 - \delta/2$
- The term (1) is more difficult to bound, we discuss this next
Uniform bound on the risk of a random $n$-dataset

$$R_n(\hat{D}) - \frac{1}{T} \sum_{t=1}^{T} \hat{R}_{\hat{D}}(z_t)$$

**Theorem 2.** With probability at least $1 - \delta/2$

$$\sup_{D \in \mathcal{D}} R_n(D) - \frac{1}{T} \sum_{t=1}^{T} \hat{R}_D(z_t) \leq$$

$$L\alpha K \sqrt{\frac{2\pi S_1(X)}{T}} + 4L\alpha \sqrt{\frac{S_{\infty}(\mathcal{E})(2 + \ln K)}{n}} + \sqrt{\frac{9 \ln 4/\delta}{2T}}$$
Uniform bound (cont.)

We decompose the uniform deviation

$$\sup_D \left\{ R_n(D) - \frac{1}{T} \sum_{t=1}^{T} \hat{R}_D(z_t) \right\}$$

as the sum of the two terms,

$$\sup_D \left\{ R_n(D) - \mathbb{E}_{z \sim \hat{\mathcal{E}}} \hat{R}_D(z) \right\} + \sup_D \left\{ \mathbb{E}_{z \sim \hat{\mathcal{E}}} \hat{R}_D(z) - \frac{1}{T} \sum_{t=1}^{T} \hat{R}_D(z_t) \right\}$$

- term $A$ is essentially the estimation error for the Lasso
- term $B$ is the estimation of the expected empirical error
Estimation bound for the Lasso

We have

\[ A = \sup_{D} \mathbb{E}_{\mu \sim \mathcal{E}} \mathbb{E}_{z \sim \mu^n} \left[ \mathbb{E}_{(x,y) \sim \mu} \ell \left( \langle D \gamma_D(z), x \rangle, y \right) - \hat{R}_D(z) \right] \]

and \( A(\mu) \) is the estimation error of the Lasso on task \( \mu \)
A useful lemma - I

Lemma 3. If $F_k = |\langle v_k, \sum_i \sigma_i x_i \rangle|$, for $\|v_k\| \leq 1$, $k = 1, \ldots, K$

$$
\mathbb{E} \max_k \left| \left\langle v_k, \sum_i \sigma_i x_i \right\rangle \right| \leq \sqrt{2n \left\| \hat{\Sigma}(x) \right\|_\infty} \left( 2 + \sqrt{\ln K} \right).
$$

Proof. Let $c = \sqrt{m \left\| \hat{\Sigma}(x) \right\|_\infty}$. By integration by parts

$$
\mathbb{E} \max F_k = \int_0^\infty \Pr \left\{ \max_k F_k \geq s \right\} ds
\leq c + \delta + \int_{c+\delta}^\infty \Pr \left\{ \max_k F_k \geq s \right\} ds
\leq c + \delta + \sum_k \int_{\delta}^\infty \Pr \{ F_k \geq \mathbb{E}F_k + s \} ds = (\star)
$$

The 2nd ineq. follows by noting that $\mathbb{E}F_k \leq c$ and a union bound
A useful lemma - II

Note that $F_k(\sigma)$ satisfies the bounded difference condition with $A^2 = 4m \left\| \hat{\Sigma}(x) \right\|_\infty$. Hence

\[
(*) \leq c + \delta + \sum_k \int_{\delta}^{\infty} \exp \left( -\frac{s^2}{2n \left\| \hat{\Sigma}(x) \right\|_\infty} \right) ds \\
\leq c + \delta + \frac{mK \left\| \hat{\Sigma}(x) \right\|_\infty}{\delta} \exp \left( -\frac{\delta^2}{2n \left\| \hat{\Sigma}(x) \right\|_\infty} \right)
\]

Finally set $\delta = \sqrt{2n \left\| \hat{\Sigma}(x) \right\|_\infty \ln (eK)}$ and compute the bound

**Note:** when computing $c$ we have maximized over $\nu_k \in H$, $\| \nu_k \| \leq 1$, however the same bound holds with $\max \langle \nu_k, \hat{\Sigma}(x)\nu_k \rangle$ in place of $\| \hat{\Sigma}(x) \|_\infty$
We can slightly improve the result and its proof using the fact that the random variable $\xi_k = \langle v_k, \sum_i \sigma_i x_i \rangle$ is subgaussian with constant $s_k^2 = m \langle v_k, \hat{\Sigma}(x)v_k \rangle$, that is, $\mathbb{E} \exp(\beta \xi_k) \leq \exp(s_k^2 \beta^2/2)$

$$
\exp \left( \beta \mathbb{E} \max_k \xi_k \right) \leq \mathbb{E} \exp \left( \beta \max_k \xi_k \right) \\
= \mathbb{E} \max_k \exp (\beta \xi_k) \\
\leq \sum_k \mathbb{E} \exp (\beta \xi_k) \\
\leq K \exp \left( \beta^2 m \max_k s_k^2 \right)
$$

Then take the log, divide by $\beta$ and optimize over $\beta$

We obtain $\mathbb{E} \max_k \xi_k \leq \sqrt{8m \max_k s_k^2 \ln K}$ and $\mathbb{E} \max_k |\xi_k|$ as above with $K$ replaced by $2K$
Bounding A

Recall $A(\mu) = \mathbb{E}_{(x,y) \sim \mu} \ell (\langle D \gamma_D(z), x \rangle, y) - \hat{R}_D(z)$

For any fixed dictionary $D$ and any measure $\mu$

$$A(\mu) \leq \mathbb{E}_{z \sim \mu^n} \sup_{\gamma \in \mathcal{C}_\alpha} \left[ \mathbb{E}_{(x,y) \sim \mu} \left[ \ell (\langle D \gamma, x \rangle, y) \right] - \frac{1}{n} \sum_i \ell (\langle D \gamma, x_i \rangle, y_i) \right]$$

$$\leq \frac{2}{n} \mathbb{E}_{z \sim \mu^n} \mathbb{E}_\sigma \sup_{\gamma \in \mathcal{C}_\alpha} \sum_{i=1}^n \sigma_i \ell (\langle D \gamma, x_i \rangle, y_i) \quad \text{[symmetrization]}$$

$$\leq \frac{2L}{n} \mathbb{E}_{z \sim \mu^n} \mathbb{E}_\sigma \sup_{\gamma \in \mathcal{C}_\alpha} \sum_k \gamma_k \langle D e_k, \sum_i \sigma_i x_i \rangle \quad \text{[contraction lemma]}$$

$$\leq \frac{2L \alpha}{n} \mathbb{E}_{z \sim \mu^n} \mathbb{E}_\sigma \max_k |\langle D e_k, \sum_i \sigma_i x_i \rangle| \quad \text{[Hölder's ineq.]}$$
Bounding A (cont.)

Continuing the last quantity is bounded as

\[
\leq \frac{2L\alpha}{n} \mathbb{E}_{z \sim \mu^n} \sqrt{2n \| \hat{\Sigma}(x) \|_\infty} \left( 2 + \sqrt{\ln K} \right) \quad \text{[Lemma 3]}
\]

\[
\leq 2L\alpha \sqrt{\frac{4\mathbb{E}_{z \sim \mu^n} \| \hat{\Sigma}(x) \|_\infty (2 + \ln K)}{n}} \quad \text{[Jensen’s ineq.]}.
\]

This gives the bound

\[
A(\mu) \leq 4L\alpha \sqrt{\frac{\mathbb{E}_{z \sim \mu^n} \| \hat{\Sigma}(x) \|_\infty (2 + \ln K)}{n}} \quad (5)
\]

valid for every measure \( \mu \) on \( H \times \mathbb{R} \) and every \( D \in \mathcal{D}_K \). Then take the expectation over \( \mu \sim \mathcal{E} \) and use Jensen’s inequality
Bounding B - I

We bound the uniform estimate of the empirical error

\[ B = \sup_{D} \left\{ \mathbb{E}_{\hat{\mathcal{E}}} \hat{R}_D(z) - \frac{1}{T} \sum_{t=1}^{T} \hat{R}_D(z_t) \right\} \]

as follows

- Using a symmetrization argument and bonding the Rademacher average by the Gaussian average we obtain with prob. \( \geq 1 - \delta \) in the multisample that

\[ B \leq \frac{\sqrt{2\pi}}{T} \mathbb{E}_{\zeta} \sup_{D \in \mathcal{D}_K} \sum_{t=1}^{T} \zeta_t \hat{R}_D(z_t) + \sqrt{\frac{9 \ln 2}{2T \delta}} \]

where \( \zeta_t \) is an orthogaussian sequence

- To bound the Gaussian average we use Slepian's Lemma
Bounding B - II

Define two Gaussian processes $\Omega$ and $\Theta$ indexed by $\mathcal{D}_K$ as

$$\Omega_D = \sum_{t=1}^{T} \zeta_t \hat{R}_D (z_t)$$

and

$$\Theta_D = \frac{L\alpha}{\sqrt{n}} \sum_{t=1}^{T} \sum_{i=1}^{n} \sum_{k=1}^{K} \zeta_{tik} \langle D_{e_k}, x_{ti} \rangle$$

where the $\zeta_{ti}$ are also orthogaussian.

Then for $D_1, D_2 \in \mathcal{D}_K$ show that

$$\mathbb{E} (\Omega_{D_1} - \Omega_{D_2})^2 = \sum_{t=1}^{T} \left( \hat{R}_{D_1} (z_t) - \hat{R}_{D_2} (z_t) \right)^2 \leq \mathbb{E} (\Theta_{D_1} - \Theta_{D_2})^2$$

Then use Slepian’s Lemma to bound

$$\mathbb{E} \sup_D \sum_t \zeta_j \hat{R}_D (z_t) \leq \frac{L\alpha}{\sqrt{n}} \mathbb{E} \sup_{D \in \mathcal{D}_K} \sum_{t,i,k} \zeta_{tik} \langle D_{e_k}, x_{ti} \rangle$$
Bounding B - III

\[
\mathbb{E} \sup_D \sum_{t,i,k} \zeta_{tik} \langle D_{ek}, x_{ti} \rangle = \mathbb{E} \sup_D \sum_k \langle D_{ek}, \sum_t \sum_i \gamma_{kij} x_{ti} \rangle
\]

\[
\leq \sup_D \sqrt{\sum_k \| D_{ek} \|^2} \mathbb{E} \sqrt{\sum_k \left\| \sum_{t,i} \zeta_{tik} x_{ti} \right\|^2}
\]

\[
\leq \sqrt{K} \sqrt{\sum_k \mathbb{E} \left\| \sum_{t,i} \zeta_{tik} x_{ti} \right\|^2}
\]

\[
\leq \sqrt{K} \sqrt{\sum_k \sum_{t,i} \| x_{ti} \|^2} \leq K \sqrt{TS_1(X)}
\]
Bounding B - IV

We therefore have that with probability at least $1 - \delta$ in the draw of the multi sample $z_1, \ldots, z_T \sim \hat{\mathcal{E}}$

$$\sup_{D \in \mathcal{D}_K} \mathbb{E}_{z \sim \hat{\mathcal{E}}} \left[ \hat{R}_D(z) \right] - \frac{1}{T} \sum_{i=1}^{T} \hat{R}_D(z_t)$$

$$\leq L \alpha K \sqrt{\frac{2 \pi S_1(X)}{T}} + \sqrt{\frac{9 \ln 2/\delta}{2T}}$$

Combing the bounds for quantities $A$ and $B$, we obtain the bound in Theorem 2

Then use this in the decomposition for $R_n(\hat{D}) - R^*$ above to obtain Theorem 1
Experiment

Learn a dictionary for image reconstruction from few pixel values (input space is the set of possible pixels indices, output space represents the gray level)

Found dictionary (top) vs. dictionary by standard SC (bottom):
Bounds for MTL

Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}^T$ and define

$$\mathcal{R}(\mathcal{F})(\mathbf{X}) = \frac{2}{nT} \mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \sum_{t=1}^{T} \sum_{i=1}^{n} \sigma_{ti} f_t(x_{ti}), \quad \mathcal{R}(\mathcal{F}) = \mathbb{E}\mathcal{R}(\mathcal{F})(\mathbf{X})$$

**Theorem 4.** Let $\delta \in (0, 1)$. With prob. $\geq 1 - \delta$ for every $f \in \mathcal{F}$

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} f_t - \frac{1}{T} \sum_{t=1}^{T} \frac{1}{n} \sum_{i=1}^{n} f_t(x_{ti}) \leq \mathcal{R}(\mathcal{F}) + \sqrt{\frac{\log \frac{1}{\delta}}{2nT}} \quad (i)$$

With the same probability

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} f_t - \frac{1}{T} \sum_{t=1}^{T} \frac{1}{n} \sum_{i=1}^{n} f_t(x_{ti}) \leq \mathcal{R}(\mathcal{F})(\mathbf{x}) + \sqrt{\frac{9 \log \frac{2}{\delta}}{2nT}} \quad (ii)$$

The proof is very similar to the proof for the scalar case [1]
Recall the multitask error and the empirical error

\[ R(f) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_t \ell(f_t(X_t), Y_t)), \quad \hat{R}(f) = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{n} \sum_{i=1}^{n} \ell(f_t(x_{ti}), y_{ti})) \]

If \( \ell(y, \cdot) \) is \( L \)-Lipschitz then the Rademacher average of the loss class is bounded as (see e.g. [3])

\[ \mathcal{R}(\ell \circ \mathcal{F}) \leq L \mathcal{R}(\mathcal{F})(\mathbf{X}) \]

so that by Theorem 4

\[ R(f) - \hat{R}(f) \leq L \mathcal{R}(\mathcal{F}) + \sqrt{\frac{\log \frac{1}{\delta}}{2nT}} \]

\[ R(f) - \hat{R}(f) \leq L \mathcal{R}(\mathcal{F})(\mathbf{X}) + \sqrt{\frac{9 \log \frac{2}{\delta}}{2nT}} \]
MTL Analysis

Goal is to bound the excess error

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \ell(\langle \hat{D}^t, x \rangle, y) - \min_{D \in \mathcal{D}_K} \frac{1}{T} \sum_{t=1}^{T} \min_{\gamma_t} \mathbb{E} \ell(\langle D^t, x \rangle, y)$$

Assumptions: $\ell(y, \cdot)$ is $L$-Lipschitz and $\|x_{ti}\| \leq 1$ a.s.

Approach: we bound the uniform deviation (use the notation $\gamma = [\gamma_1, \ldots, \gamma_T]$)

$$\sup_{D, \gamma} \left\{ \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \ell(\langle D^t, x \rangle, y) - \frac{1}{T} \sum_{t=1}^{T} \frac{1}{n} \sum_{i=1}^{n} \ell(\langle D^t, x_{ti} \rangle, y_{ti}) \right\}$$
**Bound for MTL**

**Theorem 5.** Let $\hat{S}_p := \frac{1}{T} \sum_{t=1}^{T} \|\hat{\Sigma}_t\|_p, \ p \geq 1$. With prob. $\geq 1 - \delta$ in the draw of $z_1, \ldots, z_T$ the excess error is upper bounded by

$$L_\alpha \sqrt{\frac{8S_\infty \log(2K)}{n}} + L_\alpha \sqrt{\frac{2\hat{S}_1(K+12)}{nT}} + \sqrt{\frac{8 \log \frac{4}{\delta}}{nT}}$$

- We will see that if $T$ grows, bound is **comparable to Lasso** with best a-priori known dictionary!
Proof Idea

We bound the Rademacher average of the vector-valued function class:

\[
R = \frac{2}{nT} \mathbb{E}_{\sigma} \sup_{\gamma \in C^T} \sup_{\|D\|_{2,\infty} \leq 1} \sum_{t=1}^{T} \sum_{i=1}^{n} \sigma_{ti} \langle D\gamma_t, x_{ti} \rangle
\]

\[
F_{\gamma}(\sigma)
\]

- Step 1: we show that \( F_{\gamma}(\sigma) \) is concentrated around its mean

- Step 2: we use step 1 and the fact that the supremum over each \( \gamma_t \) is attained at an extreme point of the \( \ell_1 \) ball

Once we have \( R \) the uniform bound follows from Theorem 4
Bounding $F_\gamma(\sigma)$

Lemma 6.

(i) If $\|\gamma_t\|_2 \leq 1$ for all $t$, then

$$\mathbb{E} F_\gamma \leq \sqrt{mTK} \ S_1(X)$$

(ii) If $\gamma$ satisfies $\|\gamma_t\|_1 \leq 1$ for all $t$, then for any $s \geq 0$

$$\Pr \{ F_\gamma \geq \mathbb{E} [F_\gamma] + s \} \leq \exp \left( \frac{-s^2}{8mT \ S_\infty(X)} \right)$$

The first inequality is elementary. The second uses the following concentration inequality [3, 8] which is a generalization of the bounded difference inequality.
Concentration inequality

**Theorem 7.** [8, 3] Let $F : \mathcal{X}^n \to \mathbb{R}$ and define

$$
B^2 = \sup_{x \in \mathcal{X}^n} \sum_{k=1}^{n} \left( F(x) - \inf_{y \in \mathcal{X}} F(x_k \leftarrow y) \right)^2
$$

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a vector of independent random variables with values in $\mathcal{X}$, and let $\mathbf{X}'$ be i.i.d. to $\mathbf{X}$. Then for any $s > 0$

$$
\Pr \{ F(\mathbf{X}) > \mathbb{E} F(\mathbf{X}') + s \} \leq e^{-s^2/(2B^2)}
$$

Note: the bounded difference inequality uses the stronger assumption

$$
A^2 = \sup_{x \in \mathcal{X}^n} \sum_{k=1}^{n} \sup_{y_1, y_2 \in \mathcal{X}} (F(x_k \leftarrow y_1) - F(x_k \leftarrow y_2))^2
$$

improving the tail upper bound to $e^{-2s^2/A^2}$
Proof of inequality (ii) - I

Let $D(\sigma)$ be the maximizer in $F_\gamma(\sigma) = \sup_D \sum_{t,i} \sigma_{ti} \langle D_{\gamma t}, x_{ti} \rangle$

For any $s \in \{1, \ldots, T\}$, $j \in \{1, \ldots, m\}$ and any $\sigma' \in \{-1, 1\}$

$$F_\gamma(\sigma) - F_\gamma(\sigma_{sj \leftarrow \sigma'}) \leq 2 \left| \langle D(\sigma) \gamma_s, x_{sj} \rangle \right|$$

Then

$$\sum_{sj} \left[ F_\gamma(\sigma) - \inf_{\sigma' \in \{-1,1\}} F_\gamma(\sigma_{sj \leftarrow \sigma'}) \right]^2 \leq 4 \sum_{t,i} \langle D(\sigma)\gamma_t, x_{ti} \rangle^2$$

$$\leq 4m \sum_t \left\langle D(\sigma)\gamma_t, \hat{\Sigma}(x_t)D(\sigma)\gamma_t \right\rangle$$

$$\leq 4m \sum_t \left\| \hat{\Sigma}(x_t) \right\|_\infty \left\| D(\sigma)\gamma_t \right\|^2$$

$$\leq 4m \sum_t \left\| \hat{\Sigma}(x_t) \right\|_\infty = 4m TS_\infty(X)$$

In the last ineq. we used $\|D_{\gamma t}\| \leq \sum_k |\gamma_{tk}| \|De_k\| \leq \|\gamma_t\|_1 \leq 1$
Proof of inequality (ii) - II

We first used the fact that convex functions on a compact convex set attain their maxima at an extreme point and then integration by parts and the fact that the random variable $F_\gamma \geq 0$

Let $c = \sqrt{mKTS_1(X)}$. We have

$$(*) \leq c + \delta + \sum_{\gamma \in (\text{ext}(C))^T} \int_{c+\delta}^{\infty} \Pr\{F_\gamma > s\} \, ds$$

$$\leq c + \delta + \sum_{\gamma \in (\text{ext}(C))^T} \int_{\delta}^{\infty} \Pr\{F_\gamma > E F_\gamma + s\} \, ds \quad (\text{eq. (i) Lemma 6})$$

$$\leq c + \delta + (2K)^T \int_{\delta}^{\infty} \exp \left( \frac{-s^2}{8mTS_{\infty}(X)} \right) \, ds \quad (\text{eq. (ii) and } |\text{ext}(C)| = 2K)$$

$$\leq c + \delta + \frac{4mTS_{\infty}(X)(2K)^T}{\delta} \exp \left( \frac{-\delta^2}{8mTS_{\infty}(X)} \right)$$
Proof of inequality (ii) - III

Setting \( \delta = \sqrt{8mTS_\infty(X) \ln e(2K)^T} \) we obtain

\[
(*) \leq \sqrt{2nT (K + 12) S_1(X)} + T \sqrt{8nS_\infty(X) \ln (2K)}
\]

so that

\[
R = \frac{2\alpha}{nT} A \leq 2 \sqrt{\frac{2(K + 12) S_1(X)}{nT}} + \sqrt{\frac{8S_\infty(X) \ln (2K)}{n}}
\] (**)
Using the total $\ell_1$ norm of the code vector

The original sparse coding method [13] uses the regularizer

$$\sum_{t=1}^{T} \sum_{k=1}^{K} |\gamma_{tk}|$$

If we require this to be bounded by $\alpha T$ we obtain the bound

$$\sqrt{mTKS_1(X)} + 4T \sqrt{2mS'_\infty(X) \ln(KT)}, \quad S'_\infty(X) = \max_t \|\hat{\Sigma}_t\|_\infty$$

where

$$S'_\infty(X) = \max_t \|\hat{\Sigma}(x_t)\|_\infty$$
Using the total $\ell_1$ norm of the code vector (cont.)

The proof is almost identical except that when bounding $\mathbb{E} \sup_{\gamma \in \mathcal{C}} F_\gamma$, we use the fact that the maximum is attained at one of the $2KT$ extreme points, then proceed in a very similar way and choose $\delta = \sqrt{8mT^2S'_\infty(\mathbf{X}) \ln(KT)}$

Multiplying by $2\alpha/(nT)$ we obtain the Rademacher bound

$$R \leq 2\sqrt{\frac{KS_1(\mathbf{X})}{mT}} + 8\sqrt{\frac{2S'_\infty(\mathbf{X}) \ln(KT)}{m}}$$

- The bound is larger than that for $\|\gamma_t\| \leq \alpha$ in (***) and in fact diverges for $T \to \infty$. However the empirical error can be much smaller
Subspace Learning

Same as before but now we constrain the $\ell_2$ norm on the code vectors

$$
\min_{D \in \mathcal{D}_K} \frac{1}{T} \sum_{t=1}^{T} \min_{\|\gamma\|_2 \leq \alpha} \frac{1}{n} \sum_{i=1}^{n} \ell(\langle D\gamma, x_{ti} \rangle, y_{ti})
$$

In addition we require the dictionary elements to be orthonormal, so we consider the set of matrices (we allow each column to be in Hilbert space $H$)

$$
\mathcal{W} = \{ [w_1, \ldots, w_T] : w_t = D\gamma_t, \quad D^* D = I_{K \times K}, \quad \max_t \|\gamma_t\|_2 \leq 1 \}
$$

We give a bound for the corresponding multitask Rademacher average

$$
\mathcal{R} = \frac{2}{nT} \mathbb{E}_\sigma \sup_{W \in \mathcal{W}} \sum_{t=1}^{T} \sum_{i=1}^{n} \sigma_{ti} \langle w_t, x_{ti} \rangle
$$
Theorem 8 ([12]). Let \( \hat{C} \) be the total empirical covariance. We have

\[
\mathcal{R} \leq \inf_{\eta > 0} \left\{ 2\eta \sqrt{\frac{\| \hat{C} \|_1}{n}} + 8\sqrt{n} \frac{K \| \hat{C} \|_\infty \ln \frac{4}{\eta}}{n} \right\} + 2\sqrt{\frac{K \| \hat{C} \|_1}{nT}}
\]

If \( H = \mathbb{R}^d \) we may set \( \eta = \sqrt{K/d} \) to obtain

\[
\mathcal{R} \leq 8\sqrt{\frac{K \| \hat{C} \|_\infty \ln (16 d K)}{n}} + 2\sqrt{\frac{K \| \hat{C} \|_1}{nT}}
\]
Subspace Learning

A key tool in proving the Rademacher bound is the following

**Lemma 9 ([12])**. Let \( \{\mathcal{F}\}_{m=1}^{M} \) be a collection of real-valued function classes on a set \( \mathcal{X} \), let \( x = (x_1, \ldots, x_n) \in \mathcal{X}^n \) and define

\[
Q(x) = \sup_{f \in \bigcup_m \mathcal{F}_m} \sqrt{\frac{1}{n} \sum_{i=1}^{n} f^2(x_i)}
\]

Then

\[
\mathcal{R}(\bigcup_m \mathcal{F}_m(x)) \leq \max_{m=1}^{M} \mathcal{R}(\mathcal{F}_m(x)) + 8Q(x)\sqrt{\ln M \over n}
\]
Rewriting the Rademacher average of unions

\[ \mathcal{R} \left( \bigcup_m \mathcal{F}_m(x) \right) = \frac{2}{n} \mathbb{E} \max_m \left[ \sup_{f \in \mathcal{F}_m} \sum_{i=1}^{n} \sigma_i f(x_i) \right] = \frac{2}{n} \mathbb{E} \max_m F_m(\sigma) \]

with \( F_m(\sigma) = \sup_{f \in \mathcal{F}_m} \sum_{i=1}^{n} \sigma_i f(x_i) \), for \( z \in \mathbb{R}^n \).

- \( F_m \) is convex on \( \mathbb{R}^n \)
- By Cauchy Schwarz \( F_m \) is Lipschitz with constant

\[ \sqrt{\sup_{f \in \mathcal{F}_m} \sum_{i=1}^{n} f^2(x_i)} \leq \sqrt{\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} f^2(x_i)} =: L \]
Concentration of convex Lipschitz functions

**Theorem 10 ([3]).** If $F : [-1, 1]^n \rightarrow \mathbb{R}$ is convex and $L$-Lipschitz and $X = (X_1, ..., X_n)$ is a vector of independent r.v. with values in $[-1, 1]$ then for every $\beta > 0$

$$\mathbb{E} \exp (\beta [F (X) - \mathbb{E} F (X)]) \leq e^{2\beta^2 L^2}$$
Bounding the Rademacher average of unions

\[ \forall \beta > 0, \quad \exp \left( \beta \mathbb{E} \max_m (F_m - \mathbb{E} F_m) \right) \leq \mathbb{E} \max_m \exp \left( \beta \left( F_m - \mathbb{E} F_m \right) \right) \]
\[ \leq \sum_m \mathbb{E} \exp \left( \beta \left( F_m - \mathbb{E} F_m \right) \right) \]
\[ \leq M e^{2\beta^2 L^2} \]

Take log, divide by \( \beta \), use triangle inequality and optimize in \( \beta \)

\[ \mathbb{E} \max_m (F_m) \leq \max_m \mathbb{E} F_m + \frac{\ln M}{\beta} + 2\beta L^2 \]
\[ \leq \max_m \mathbb{E} F_m + L \sqrt{8 \ln M} \]

We have shown:

\[ \mathcal{R} \left( \bigcup_m \mathcal{F}_m(x) \right) \leq \max_m \mathcal{R} (\mathcal{F}_m(x)) + \frac{8}{n} \sqrt{\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} f^2(x_i) \ln M} \]
Subspace Learning

We first construct a finite approximation of $\Gamma = \{\max_t \|\gamma_t\|_2 \leq 1\}$

- for every $\eta > 0$ there exists a finite subset $\Gamma_0 \subset \Gamma$ such that
  \[ \forall \gamma \in \Gamma, \exists \beta \in \Gamma_0 \text{ such that } \|\gamma - \beta\| \leq \eta \text{ and } |\Gamma_0| \leq (4/\eta)^{KT} \]

- For every $\beta \in \Gamma_0$ let $\Gamma_\beta = \{\gamma \in \Gamma : \|\gamma - \beta\| \leq \eta\}$, so that
  \[ \Gamma = \bigcup_{\beta \in \Gamma_0} \Gamma_\beta \]

Hence

\[ \mathcal{W} = \bigcup_{\beta \in \Gamma_0} \{W = D\gamma : D^*D = I_{K \times K}\} \]

To prove Theorem 8 we apply Lemma 9 to the induced finite union of linear vector-valued functions
Subspace Learning

By orthonormality of the dictionary

\[
\max_{\gamma_t} \sup_{D^* D = I} \frac{1}{nT} \sum_{t,i} \left( \sum_k \gamma_{tk} d_k, x_{ti} \right)^2 = \max_{\|v\| \leq 1} \frac{1}{nT} \sum_{t,i} \langle v, x_{ti} \rangle^2 = \lambda_{\text{max}} \left( \hat{\mathcal{C}} (x) \right)
\]

Next we bound

\[
\max_m R_m(X) = \frac{2}{nT} \max_{\beta \in \Gamma_0} \mathbb{E} \sup_{\gamma \in \Gamma_\beta} \sup_{D^* D = I} \sum_{t,i} \sigma_{ti} \sum_k \gamma_{tk} \langle d_k, x_{ti} \rangle
\]

\[
\leq \frac{2}{nT} \max_{\beta \in \Gamma_0} \mathbb{E} \sup_{D^* D = I} \sum_{t,i} \sigma_{ti} \sum_k \beta_{tk} \langle d_k, x_{ti} \rangle
\]

\[
+ \frac{2}{nT} \mathbb{E} \sup_{\|\gamma_t\|_1 < \eta} \sup_{D^* D = I} \sum_{t,i} \sigma_{ti} \sum_k \gamma_{tk} \langle d_k, x_{ti} \rangle
\]

The first term is bounded by \(2 \sqrt{K \text{ tr} \left( \hat{\mathcal{C}} (x) \right) / (nT)}\)
Subspace Learning

For the second term we again use orthonormality of the dictionary

\[
\frac{2}{nT} \mathbb{E} \sup_{\|\gamma\| < \eta} \sup_{D \in \mathcal{D}} \sum_t \left\langle \sum_k \gamma_{tk} d_k, \sum_i \sigma_{ti} x_{ti} \right\rangle
\]

\[
= \frac{2}{nT} \mathbb{E} \sup_{D \in \mathcal{D}} \sum_t \sup_{\|w\| \leq \eta} \left\langle \sum_k \gamma_k d_k, \sum_i \sigma_{ti} x_{ti} \right\rangle = \frac{2\eta}{nT} \sum_t \mathbb{E} \left\| \sum_i \sigma_{ti} x_{ti} \right\|
\]

\[
\leq \frac{2\eta}{T} \sum_t \sqrt{\frac{\text{tr} \left( \hat{C}(x_t) \right)}{n}} \leq 2\eta \sqrt{\frac{1}{T} \sum_t \text{tr} \left( \hat{C}(x_t) \right)} = 2\eta \sqrt{\frac{\text{tr} \left( \hat{C}(x) \right)}{n}},
\]

where we used Jensen’s inequality. Putting everything together and taking the infimum over \( \eta \) we get Theorem 8.
References


