AD machine learning Project

From Bound Majorization to Stochastic Bound Majorization

Yunian Pan

ECE department

May. 7th 2019

From Bound Majorization to Stochastic Bound Majorization

- Theoretical development
- Convergence Guarantee
- Evaluations
- Conclusion

Outperforming state-of-the-art first- and second-order optimization methods on various learning tasks

For a given i.i.d. dataset $\{(x_1, y_1), \ldots, (x_t, y_t)\}$, $y \in \Omega$ where $|\Omega| = K$, setting linear predictors for every data point:

$$\ln \Pr(y_i | y_i = 1, x_i) = \theta_1^\top \cdot x_i - \ln Z$$
$$\ln \Pr(y_i | y_i = 2, x_i) = \theta_2^\top \cdot x_i - \ln Z$$
$$\ldots\ldots\ldots$$
$$\ln \Pr(y_i | y_i = K, x_i) = \theta_K^\top \cdot x_i - \ln Z$$

For a given i.i.d. dataset $\{(x_1, y_1), \ldots, (x_t, y_t)\}$, $y \in \Omega$ where $|\Omega| = K$, setting linear predictors for every data point:

$$\ln \Pr(y_i | y_i = 1, x_i) = \theta_1^\top \cdot x_i - \ln Z$$
$$\ln \Pr(y_i | y_i = 2, x_i) = \theta_2^\top \cdot x_i - \ln Z$$
$$\cdots\cdots\cdots$$
$$\ln \Pr(y_i | y_i = K, x_i) = \theta_K^\top \cdot x_i - \ln Z$$

normalizer Z, prior $\Pr(y = k) = h(y)$, score $\theta_k^\top x_i \to \theta^\top \mathbf{f}_{x_i}(y)$
Resulting soft-max partition function:

$$Z_{x_i}(\theta) = \sum_{y \in \Omega} h(y) \exp(\theta^\top \mathbf{f}_{x_i}(y)) \tag{1}$$

Upper bound of Partition

Notation setting:

1. $\pi(\cdot) : \Omega \to \{1, \ldots, n\}$ s.t. $h(y) = h(\pi^{-1}(j)) = h_j$ and $\mathbf{f}(y) = \mathbf{f}(\pi^{-1}(j)) = \mathbf{f}_j$
2. $\lambda = \theta - \tilde{\theta}$
3. $Z(\theta) = \sum_{j=1}^{n} \alpha_j \exp(\lambda^\top \mathbf{f}_j)$, where $\alpha_j = h(j) \exp(\tilde{\theta}^\top \mathbf{f}_j)$.

Upper bound of Partition

Notation setting:

1. $\pi(\cdot) : \Omega \to \{1, \ldots, n\}$ s.t. $h(y) = h(\pi^{-1}(j)) = h_j$ and $\mathbf{f}(y) = \mathbf{f}(\pi^{-1}(j)) = \mathbf{f}_j$
2. $\lambda = \theta - \tilde{\theta}$
3. $Z(\theta) = \sum_{j=1}^{n} \alpha_j \exp(\lambda^\top \mathbf{f}_j)$, where $\alpha_j = h(j) \exp(\tilde{\theta}^\top \mathbf{f}_j)$.

In order to construct the monotonicity we denote $Z_i(\theta) = \sum_{j=1}^{i} \alpha_j \exp(\lambda^\top \mathbf{f}_j)$, and a trivial bound holds for $i = 0$:

$$Z_0(\theta) = 0 \leq z_0 \exp\left(\frac{1}{2}\lambda^\top \Sigma_0 \lambda + \lambda^\top \mu_0\right)$$

Where $z_0 = 0^+, \ \mu_0 = \mathbf{0}, \ \Sigma_0 = z\mathbf{I}$.

Construct bound

As we add another term $\alpha_1 \exp(\lambda^\top \mathbf{f_1})$, on both side of the above inequality, the bound still holds,

$$Z_1(\theta) \leq z_0 \exp(\frac{1}{2}\lambda^\top \Sigma_0 \lambda + \lambda^\top \mu_0) + \alpha_1 \exp(\lambda^\top \mathbf{f_1})$$

Construct bound

As we add another term $\alpha_1 \exp(\lambda^\top \mathbf{f_1})$, on both side of the above inequality, the bound still holds,

$$Z_1(\theta) \leq z_0 \exp(\frac{1}{2}\lambda^\top \Sigma_0 \lambda + \lambda^\top \mu_0) + \alpha_1 \exp(\lambda^\top \mathbf{f_1})$$

Goal: Transform the RHS into quadratic form.

$$Z_1(\theta) \leq z_1 \exp(\frac{1}{2}\lambda^\top \Sigma_1 \lambda + \lambda^\top \mu_1)$$

Construct bound

As we add another term $\alpha_1 \exp(\lambda^\top \mathbf{f_1})$, on both side of the above inequality, the bound still holds,

$$Z_1(\theta) \leq z_0 \exp(\frac{1}{2}\lambda^\top \Sigma_0 \lambda + \lambda^\top \mu_0) + \alpha_1 \exp(\lambda^\top \mathbf{f_1})$$

Goal: Transform the RHS into quadratic form.

$$Z_1(\theta) \leq z_1 \exp(\frac{1}{2}\lambda^\top \Sigma_1 \lambda + \lambda^\top \mu_1)$$

Same recurssive procedure for $Z_2(\theta), \ldots, Z_n(\theta)$.

Algebra Work

Logarithmic transformation:

$$
\begin{aligned}
\log Z_1(\theta) &\leq \log z_0 + \log(\exp(\frac{1}{2}\lambda^\top \Sigma_0 \lambda + \lambda^\top \mu_0) + \frac{\alpha_1}{z_0}\exp(\lambda^\top \mathbf{f}_1)) \\
&= \log z_0 + \log(\exp(\frac{1}{2}\lambda^\top \Sigma_0 \lambda + \lambda^\top(\mu_0 - \mathbf{f}_1)) + \frac{\alpha_1}{z_0}) + \lambda^\top \mathbf{f}_1
\end{aligned}
$$

Algebra Work

Logarithmic transformation:

$$\begin{aligned}\log Z_1(\theta) &\leq \log z_0 + \log(\exp(\frac{1}{2}\lambda^\top \Sigma_0 \lambda + \lambda^\top \mu_0) + \frac{\alpha_1}{z_0}\exp(\lambda^\top \mathbf{f}_1))\\ &= \log z_0 + \log(\exp(\frac{1}{2}\lambda^\top \Sigma_0 \lambda + \lambda^\top(\mu_0 - \mathbf{f}_1)) + \frac{\alpha_1}{z_0}) + \lambda^\top \mathbf{f}_1\end{aligned}$$

seperate $\frac{1}{2}w^\top w = \frac{1}{2}(\mathbf{f}_1 - \mu_0)^\top \Sigma_0^{-1}(\mathbf{f}_1 - \mu_0)$

$$\begin{aligned}RHS &= \log z_0 + \lambda^\top \mathbf{f}_1 - \frac{1}{2}w^\top w + \log\exp\frac{1}{2}w^\top w \cdot \exp(\frac{1}{2}\lambda^\top \Sigma_0 \lambda + \lambda^\top \mu_0) + \frac{\alpha_1}{z_0}\\ &= \log z_0 + \lambda^\top \mathbf{f}_1 - \frac{1}{2}w^\top w + \log(\exp(\frac{1}{2}u^\top u) + \gamma)\end{aligned}$$

Where $u^\top u = \frac{1}{2}(\mathbf{f}_1 - \mu_0)^\top \Sigma_0^{-1}(\mathbf{f}_1 - \mu_0) + \frac{1}{2}\lambda^\top \Sigma_0 \lambda + \lambda^\top \mu_0$, and $\gamma = \frac{\alpha}{z_0}\exp(\frac{1}{2}w^\top w)$

Useful Machinery

Lemma

For all $u \in R^d$ and $v \in R^d$ and any $\gamma \geq 0$, the bound
$\log(\exp(\frac{1}{2}\|u\|^2) + \gamma) \leq$

$$\log(\exp(\frac{1}{2}\|v\|^2)+\gamma)+\frac{v^\top(u-v)}{1+\gamma\exp(-\frac{1}{2}\|v\|^2)}+\frac{1}{2}(u-v)^\top(I+\Gamma vv^\top)(u-v)$$

holds when the scalar term $\Gamma = \frac{\tanh(\frac{1}{2}\log(\gamma\exp(-\frac{1}{2}\|v\|^2)))}{2\log(\gamma\exp(-\frac{1}{2}\|v\|^2))}$, *equality is achieved when $u = v$.*

Proof.

see T. Jebara. *Multitask sparsity via maximum entropy discrimination. JMLR, 12:75110,* 2011. □

Applying Lemma

$$\log Z_1(\theta) \leq \log z_0 + \lambda^\top \mathbf{f}_1 - \frac{1}{2}(\mathbf{f}_1 - \mu_0)^\top \Sigma_0^{-1}(\mathbf{f}_1 - \mu_0)$$
$$+ \log(\exp(\frac{1}{2}\|v\|^2) + \gamma) +$$
$$\frac{v^\top(u-v)}{1+\gamma\exp(-\frac{1}{2}\|v\|^2)} + \frac{1}{2}(u-v)^\top(I + \Gamma v v^\top)(u-v)$$

Use undetermined coefficients method, recall the goal 2:

$$z_1 = z_0 + \alpha_1$$
$$\mu_1 = \mu_0 + \frac{\alpha_1}{z_0 + \alpha_1}(\mathbf{f}_1 - \mu_0)$$
$$\Sigma_1 = \Sigma_0 + \frac{\tanh(\frac{1}{2}\log(\frac{\alpha_1}{z_0}))}{2\log(\frac{\alpha_1}{z_0})}(\mathbf{f}_1 - \mu_0)(\mathbf{f}_1 - \mu_0)^\top$$

Algorithm1

Algorithm 1: Compute Bound

Input: Parameters $\tilde{\theta}, \ \mathbf{f}(y), \ h(y)$

Initialize: $z \leftarrow 0^+, \ \mu \leftarrow 0, \ \Sigma \leftarrow zI$;

for *each* $y \in \Omega$ **do**

$\alpha = h(y)\exp(\tilde{\theta}^\top f(y))$

$\mu = \mu + \frac{\alpha}{z+\alpha}(\mathbf{f}(y) - \mu)$

$\Sigma = \Sigma + \frac{\tanh(\frac{1}{2}\log(\frac{\alpha}{z}))}{2\log(\frac{\alpha}{z})}(\mathbf{f}(y) - \mu)(\mathbf{f}(y) - \mu)^\top$

$z = z + \alpha$

end

Output: z, μ, Σ

Going back to the loglikelihood of multi-class logistic regression:

$$J(\theta) = \sum_{i=1}^{t} \left[\log \frac{h_{x_i}(y_i)}{Z_{x_i}(\theta)} + \theta^\top \mathbf{f}_{x_i}(y_i) - \frac{\lambda}{2} \|\theta\|^2 \right] \tag{2}$$

Going back to the loglikelihood of multi-class logistic regression:

$$J(\theta) = \sum_{i=1}^{t}[\log \frac{h_{x_i}(y_i)}{Z_{x_i}(\theta)} + \theta^\top \mathbf{f}_{x_i}(y_i) - \frac{\lambda}{2} \|\theta\|^2] \tag{2}$$

As we drop the terms unrelated to θ, the maximization problem becomes $\arg\min_\theta Q(\theta, \tilde{\theta})$:

$$Q(\theta, \tilde{\theta}) = \frac{1}{2}(\theta - \tilde{\theta})^\top (\textstyle\sum_i (\Sigma_i + \lambda I))(\theta - \tilde{\theta}) + \textstyle\sum_i \theta^\top (\mu_i - \mathbf{f}_{x_i}(y_i) + \lambda \tilde{\theta}) - \text{const}$$

Bound Majorization

Algorithm 2: BM

Input: Input x_i, y_i and functions $h_{x_i}, \mathbf{f}_{x_i}$ for $i = 1, 1, \ldots, t$, regularizer $\lambda \in R^+$ and convex hull $\Lambda \subseteq R^d$, tolerance ϵ

Initialize: θ_0 anywhere inside Λ and set $\tilde{\theta} = \theta_0$;

while $\theta_{new} - \theta_{old} \geq \epsilon$ **do**

  for $i = 1, \ldots, t$ **do**

    Get μ_i, Σ_i, from h_{x_i}, $\mathbf{f}_{x_i}, \tilde{\theta}$ via Algorithm 1

  end

  Set $\tilde{\theta} =$
  $\arg\min_\theta \frac{1}{2}(\theta - \tilde{\theta})^\top(\sum_i \Sigma_i + \lambda I)(\theta - \tilde{\theta}) + \theta^\top(\sum_i \mu_i - \mathbf{f}_{x_i}(y_i) + \lambda\tilde{\theta})$

  Which means: $\tilde{\theta} = \tilde{\theta} - (\sum_i \Sigma + \lambda I)^{-1}(\sum_i \mu_i - \mathbf{f}_{x_i}(y_i) + \lambda\tilde{\theta})$

end

Output: $\hat{\theta} = \tilde{\theta}$

Algorithm 3: Stochastic Bound Majorization

Input: prior $h(\cdot)$, function $\mathbf{f}(\cdot)$, regularizer $\lambda \in R^+$ and convex hull $\Lambda \subseteq R^d$ ϵ

Initialize: θ_0 anywhere inside Λ and set $\tilde{\theta} = \theta_0$;

while $\theta_{new} - \theta_{old} \geq \epsilon$ **do**

  randomly select p mini-batch x_i, y_i's

  for $i = 1, \ldots, p$ **do**

    Get $\mu_i,\ \Sigma_i$, from $h_{x_i},\ \mathbf{f}_{x_i}, \tilde{\theta}$ via Algorithm 1

  end

  Set

  $\tilde{\theta} = \arg\min_\theta \frac{1}{2}(\theta - \tilde{\theta})^\top(\sum_i \Sigma_i + \lambda I)(\theta - \tilde{\theta}) + \theta^\top(\sum_i \mu_i - \mathbf{f}_{x_i}(y_i) + \lambda\tilde{\theta})$

  Which means: $\tilde{\theta} = \tilde{\theta} - (\sum_i \Sigma + \lambda I)^{-1}(\sum_i \mu_i - \mathbf{f}_{x_i}(y_i) + \lambda\tilde{\theta})$

end

Output: $\hat{\theta} = \tilde{\theta}$

Can we do better? Yes.

Notice a linear system

$$\Big(\sum_j \Sigma_j(\theta_{n-1}) + \lambda I\Big)(\theta_n - \theta_{n-1}) = \sum_j \mu_j(\theta_{n-1}) \tag{3}$$

Applying Sherman-Morrison formula:

$\big(\Sigma + (\sqrt{\beta}I)^\top(\sqrt{\beta}I)\big)^{-1} = \Sigma^{-1} - \frac{\Sigma^{-1}(\sqrt{\beta}I)^\top(\sqrt{\beta}I)\Sigma^{-1}}{1+(\sqrt{\beta}I)^\top\Sigma^{-1}(\sqrt{\beta}I)}$,

$$M_{n+1} = M_n - \frac{\beta M_n I^\top I M_n}{1 + \beta I^\top M_n I} \tag{4}$$

Algorithm 4: SBM

Input: $h(\cdot)$, $\mathbf{f}(\cdot)$, $\lambda \in R^+$, $\Lambda \subseteq R^d$, η, ϵ

Initialize: $\theta_0 \in \Lambda$ and set $\tilde{\theta} = \theta_0$, $\phi = \mathbf{0}$, $M = \frac{1}{\lambda} I$, $\mu = 0$;

while $\theta_{new} - \theta_{old} \geq \epsilon$ **do**

  randomly select p mini-batch x_i, y_i's

  for $i = 1, \dots, p$ **do**

    $z \leftarrow 0^+$; $g = 0$

    for *each* $y \in \Omega$ **do**

      $\alpha = h(y)\exp(\tilde{\theta}^\top f(y)) \quad l = f(y) - g \quad \beta = \frac{\tanh(\frac{1}{2}\log(\frac{\alpha}{z}))}{2\log(\frac{\alpha}{z})}$

      $z = z + \alpha \quad \kappa = \frac{\alpha}{z}$

      $M = M - \frac{\beta M l^\top l M}{1 + \beta l^\top M l}$

      $\phi = \phi + M(\kappa l - \mathrm{f}_{x_i}(y) + \frac{\lambda \tilde{\theta}}{t}) - \frac{\beta M l^\top l M}{1 + \beta l^\top M l}\mu$

      $\mu = \mu + \kappa l - \mathrm{f}_{x_i}(y) + \frac{\lambda \tilde{\theta}}{t}$

      $g = g + \kappa l$

    end

  end

  $\tilde{\theta} = \tilde{\theta} - \eta\phi$

end

Output: $\hat{\theta} = \tilde{\theta}$

Define mapping: $G(\theta) := \theta - \eta V(\theta)$ where $V(\theta) = \Sigma^{-1}(\theta)\mu(\theta)$,
Fixed point equation: $\theta^* = G(\theta^*)$,
which simply indicates: $\Sigma^{-1}(\theta^*)\mu(\theta^*) = 0$.

Lemma

Define a mapping $L(\theta) := \theta - \eta V(\theta^*)$ which is equivalent to applying gradient operator $T(\theta) := \theta - \eta\nabla Q(\theta|\theta^*)$ z_θ times, i.e. $L(\theta) = T^{z_\theta}(\theta)$, where z_θ is a finte integer, and $\nabla Q(\theta|\theta^*)$ is the gradient w.r.t population, under strong convexity condition and smoothness assumption which already hold with stepsize $\eta = \frac{2}{\epsilon + l}$, and because $T(\theta)$ is contractive, we have:

$$\|L(\theta) - \theta^*\|_2 \leq \left(\frac{l - \epsilon}{l + \epsilon}\right)^{z_\theta} \|\theta - \theta^*\|_2 \tag{5}$$

Proof.

To prove the lemma 2, leverage several truths:

- The standard result $\|T(\theta) - \theta^*\|_2 \leq \left(\frac{l-\epsilon}{l+\epsilon}\right) \|\theta - \theta^*\|_2$
- z_θ is the number of iteration that we perform to optimize a quadratic problem which is theoretically finite.
- $T^{z_\theta}(\theta_{z_\theta}) = TT^{z_\theta - 1}(\theta_{z_\theta - 1})$

Follows the inequality 8 □

Let's introduce this useful assumption analogous to gradient stability.

Definition

$V(\theta)$ stability

The functions $\{Q(\cdot|\theta), \theta \in \Omega\}$ statisfy VS(γ) condition, where $\gamma \geq 0$, over Euclidean ball $B_2(d, \theta^*)$, if

$$\left\| \Sigma(\theta)^{-1}\mu(\theta) - \Sigma(\theta^*)^{-1}\mu(\theta^*) \right\|_2 \leq \gamma \left\| \theta - \theta^* \right\|_2 \tag{6}$$

for all $\theta \in B_2(d, \theta^*)$

Back to the update:

$$
\begin{aligned}
\|G(\theta) - \theta^*\|_2 &= \|\theta - \eta V(\theta) - \theta^*\|_2 \\
&\leq \|\theta - \eta V(\theta^*) - \theta^*\|_2 + \eta \|V(\theta) - V(\theta^*)\|_2 \\
&= \|L(\theta) - \theta^*\|_2 + \eta \|V(\theta) - V(\theta^*)\|_2
\end{aligned}
$$

$$
\|G(\theta) - \theta^*\|_2 \leq \left(\left(\frac{I - \epsilon}{I + \epsilon} \right)^{z(\theta)} + \eta\gamma \right) \|\theta - \theta^*\|_2 \tag{7}
$$

the term $(\frac{I-\epsilon}{I+\epsilon})^{z(\theta)} + \eta\gamma < 1$ under a loose condition $\epsilon > \gamma$, resulting in the convergence of Bound Algorithm.

Theorem

For any $\theta_0 \in \Lambda$ all $\|\mathbf{f}_{x_i}(y)\| \leq r$ and all $|\Omega| \leq n$, Algorithm 2 outputs a θ s.t. $J(\theta_\tau) - J(\theta_\tau) \leq \epsilon(J(\theta^*) - J(\theta_\tau))$ with more than $\tau = \left\lceil \frac{\log(\epsilon)}{\log(\kappa-1) - \log\kappa} \right\rceil$ epochs of training. $\kappa = \frac{w+\lambda}{\lambda}$, and upper bound of Σ is $\omega I = (2r^2 \sum_{i=2}^{n} \frac{\tanh(\frac{1}{2}\log i)}{\log i})I$.

Proof.

See Jebara, Tony, and Anna Choromanska. "*Majorization for CRFs and latent likelihoods.*" *Advances in Neural Information Processing Systems.* 2012.

□

This is a measure of how far we have to go to achieve some accuracy.

Lemma

Define a mapping $L(\theta) := \theta - \eta V(\theta^*)$ which is equivalent to appling gradient operator $T(\theta) := \theta - \eta \nabla Q(\theta|\theta^*)$ z_θ times, i.e. $L(\theta) = T^{z_\theta}(\theta)$, where z_θ is a finte integer, and $\nabla Q(\theta|\theta^*)$ is the gradient w.r.t population, under strong convexity condition and smoothness assumption which already hold with stepsize $0 \le \eta \le \frac{2}{\epsilon+l}$, and because $T(\theta)$ is contractive, we have:

$$\|L(\theta) - \theta^*\|_2 \le (1 - \frac{2\eta l\epsilon}{l+\epsilon})^{z_\theta} \|\theta - \theta^*\|_2 \tag{8}$$

Similarly using the exactly the same technique as before we can get:

$$\|G(\theta) - \theta^*\|_2 \le ((1 - \frac{2\eta l\epsilon}{l+\epsilon})^{z_\theta} + \eta\gamma) \|\theta - \theta^*\|_2 \tag{9}$$

Denote $\Delta_{t+1} := \theta_{t+1} - \theta^*$, we have that:

$$\left\|\Delta_{t+1}\right\|_2^2 - \left\|\Delta_t\right\|_2^2 \leq (\eta_t)^2 \left\|\hat{V}(\theta_t)\right\|_2^2 + 2\eta_t \left\|\hat{V}(\theta_t) \cdot \Delta_t\right\|_2$$
$$\implies E[\left\|\Delta_{t+1}\right\|_2^2] \leq E[\left\|\Delta_t\right\|_2^2] + (\eta_t)^2 E[\left\|\hat{V}(\theta_t)\right\|_2^2] + 2\eta_t E[\left\|\hat{V}(\theta_t) \cdot \Delta_t\right\|_2]$$

Since $\hat{V}(\theta^*) = 0$, we have:$E[\left\|\Delta_{t+1}\right\|_2^2] \leq$
$E[\left\|\Delta_t\right\|_2^2] + (\eta_t)^2 E[\left\|\hat{V}(\theta_t)\right\|_2^2] + 2\eta_t E[\left\|(\hat{V}(\theta_t) - \hat{V}(\theta^*)) \cdot \Delta_t\right\|_2]$
Then we upper bound the last term using
$(\left\|G(\theta) - \theta^*\right\|_2 \leq (1 - \frac{2\eta l\epsilon}{l+\epsilon})^{z_\theta} + \eta\gamma) \left\|\theta - \theta^*\right\|_2$, which is:
$2\eta_t E[\left\|(\hat{V}(\theta_t) - \hat{V}(\theta^*)) \cdot \Delta_t\right\|_2 \leq (1 - \frac{2\eta l\epsilon}{l+\epsilon})^{z_\theta} + \eta\gamma - 1) \left\|\theta_t - \theta^*\right\|_2$
and we get:

$$E[\left\|\Delta_{t+1}\right\|_2^2] \leq E[\left\|\Delta_t\right\|_2^2] + (\eta_t)^2 E[\left\|\hat{V}(\theta_t)\right\|_2^2]$$
$$- 2((1 - \frac{2\eta_t l\epsilon}{l + \epsilon})^{z_\theta} + \eta_t\gamma - 1) E[\left\|\Delta_t\right\|_2^2]$$

For simplicity it's safe to set $z(\theta) = 1$ as the inequality still holds and we get:

$$E[\|\Delta_{t+1}\|_2^2] \leq E[\|\Delta_t\|_2^2] + (\eta_t)^2 E[\|\hat{V}(\theta_t)\|_2^2] - 2\eta_t \xi E[\|\Delta_t\|_2^2]$$

where $\xi = \frac{2l\epsilon}{l+\epsilon} - \gamma$, combining all the previous results and upper bounding the second term $\sup_{\theta \in \Lambda} E[\|\hat{V}(\theta_t)\|_2^2] = \sigma_V^2$:

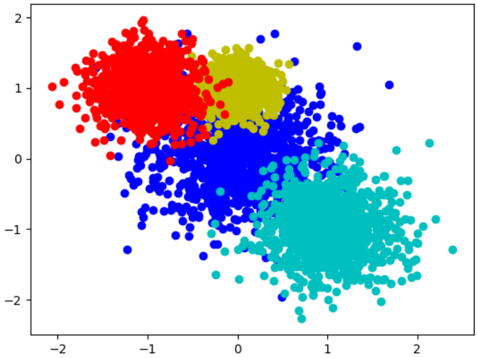
$$\begin{aligned} E[\|\Delta_{t+1}\|_2^2] &\leq (1 - 2\eta_t \xi) E[\|\Delta_t\|_2^2] + (\eta_t)^2 E[\|\hat{V}(\theta_t)\|_2^2] \\ &\leq (1 - \eta_t \xi) E[\|\Delta_t\|_2^2] + (\eta_t)^2 \sigma_V^2 \end{aligned}$$

Setting $\eta_t = \frac{3}{2\xi(t+2)}$ and unwrapping the recursion, after some algebra work including summation, multiplication, contraction and upper bounding,

$$E[\|\Delta_{t+1}\|_2^2] \leq \frac{9\sigma_V^2}{\xi^2}\frac{1}{t+2} + (\frac{2}{t+2})^{\frac{3}{2}} \|\Delta_0\|_2^2 \tag{10}$$

Which summarize the guarantee of convergence.

$t = 4000$ and $n = 4$, We simply choose $h(y) = \frac{\mathbb{1}(y=k)}{\sum_{k=1}^{4} \mathbb{1}(y=k)}$ to be the prior, and
$f_x(y) = [\mathbb{1}(y = 1)x^\top \mathbb{1}(y = 2)x^\top, \mathbb{1}(y = 3)x^\top, \mathbb{1}(y = 4)x^\top]^\top$ to be the mapping.



Explored SBM, BM, LBFGS, GD and SGD whose parameter settings are tuned and shown in table 1

Table 1: parameter setting

p : batch size		m : number of vectors in LBFGS		
BM	SBM	LBFGS	GD	SGD
$\lambda = 1e-2$	$p = 40$	η : line search	$\lambda = 1e-2$	$p = 40$
$\epsilon = 1e-6$	$\epsilon = 1e-6$	$\epsilon = 1e-5$	$\epsilon = 1e-5$	$\epsilon = 1e-5$
$\eta = 1$	$\eta = 1e-2$	$m = 4$	$\eta = 1e-2$	$\eta = 1e-2$

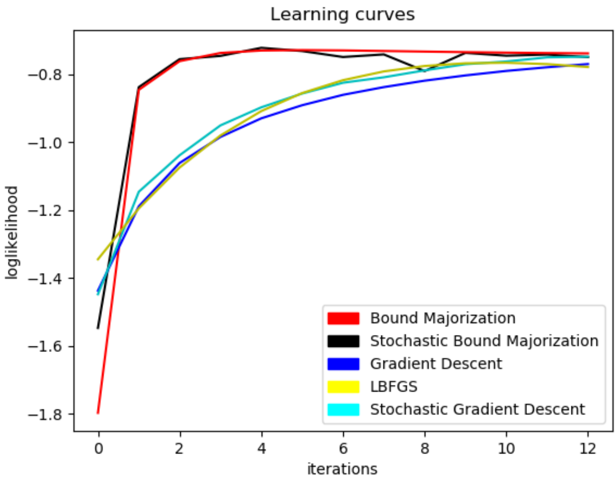


Figure 1: iteration comparison

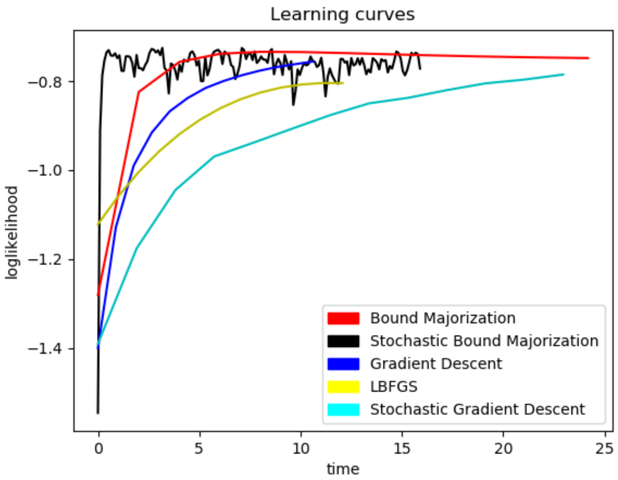


Figure 2: time comparison

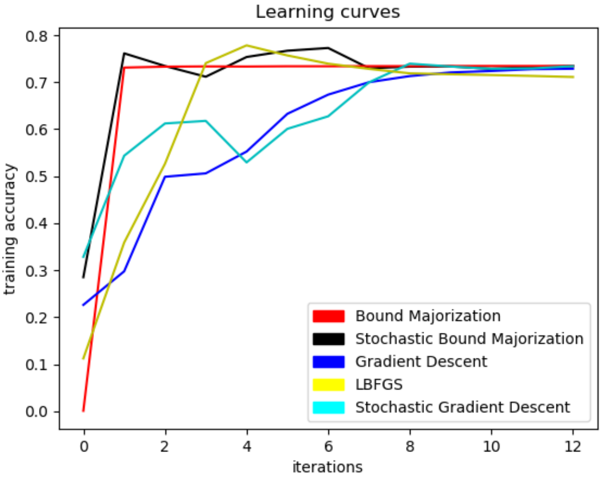


Figure 3: training accuracy comparison

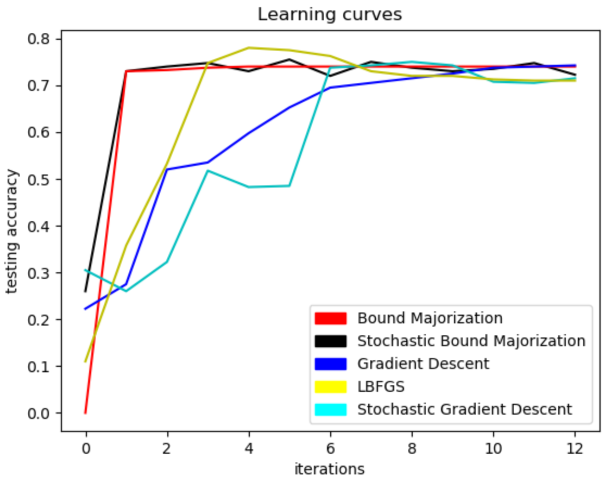


Figure 4: testing accuracy comparison

Conclusion

- Requiring very few parameters tuning, (stepsize η or convex hull Λ);
- Bound is very tight, which makes it extremely efficient;
- Only applicable to log-linear models, CRFs, Latent Likelihoods etc.
- The assumptions and conditions has to be satisfied properly, otherwise it may diverge.

Thank You!