Convergence of Heterogeneous Distributed Learning In Stochastic Routing Game

Syrine Krichene  Walid Krichene  Roy Dong  Alexandre Bayen

September 30, 2015
Outline

1 Introduction

2 Heterogeneous Learning with Stochastic Mirror Descent

3 Simulations

References
Routing game

Used to model congestion in
- Transportation networks
- Communication networks
Routing game

Used to model congestion in

- Transportation networks
- Communication networks

![Directed graph (V, E)](image)

**Figure**: Example network

- Directed graph \((V, E)\)
- Population \(k\): paths \(P_k\)
Routing game

Used to model congestion in
- Transportation networks
- Communication networks

![Directed graph \((V, E)\)](image)

**Figure:** Example network

- Directed graph \((V, E)\)
- Population \(k\): paths \(\mathcal{P}_k\)
- Population distribution over paths \(x_{\mathcal{P}_k} \in \Delta^{\mathcal{P}_k}\)
- Loss on path \(p\): 
  \[
  \ell_p(x) = \sum_{e \in p} c_e(\phi_e)
  \]
Routing game

Used to model congestion in

- Transportation networks
- Communication networks

Directed graph \((V, E)\)
- Population \(k\): paths \(P_k\)
- Population distribution over paths \(x_{P_k} \in \Delta P_k\)
- Loss on path \(p\): \(\ell_p(x) = \sum_{e \in p} c_e(\phi_e)\)
Online learning model

Online Learning Model

1: for $t \in \mathbb{N}$ do
2: \hspace{1em} Play $p \sim x_{\mathcal{P}_k}^{(t)}$
3: \hspace{1em} Discover $\ell_{\mathcal{P}_k}^{(t)}$
4: \hspace{1em} Update $x_{\mathcal{P}_k}^{(t+1)}$
5: end for

$x_{\mathcal{P}_1}^{(t)} \in \Delta_{\mathcal{P}_1}$

Sample $p \sim x_{\mathcal{P}_k}^{(t)}$
Discover $\ell_{\mathcal{P}_k}^{(t)}$
Update $x_{\mathcal{P}_k}^{(t+1)}$
Online learning model

Online Learning Model

1: for $t \in \mathbb{N}$ do
2: Play $p \sim x_{P_k}^{(t)}$
3: Discover $\ell_{P_k}^{(t)}$
4: Update $x_{P_k}^{(t+1)}$
5: end for
Online learning model

Online Learning Model

1: for $t \in \mathbb{N}$ do
2:   Play $p \sim x_{P_k}^{(t)}$
3:   Discover $\ell_{P_k}^{(t)}$
4:   Update $x_{P_k}^{(t+1)}$
5: end for
Online learning model

---

**Online Learning Model**

1. \textbf{for} \( t \in \mathbb{N} \) \textbf{do}
2. \hspace{1em} Play \( p \sim x_{P_1}^{(t)} \)
3. \hspace{1em} Discover \( \ell_{P_1}^{(t)} \)
4. \hspace{1em} Update \( x_{P_1}^{(t+1)} \)
5. \textbf{end for}

---

![Online learning model diagram](image)
Convergence to Nash equilibria

**Nash equilibrium**

$x^*$ is a Nash equilibrium if for all $x$

$$
\langle \ell(x^*), x - x^* \rangle = \sum_k \langle \ell_{P_k}(x^*), x_{P_k} - x_{P_k}^* \rangle \geq 0
$$

I.e., for each population, every path in the support of $x_{P_k}^*$ has minimal loss.
Convergence to Nash equilibria

Nash equilibrium

\( x^* \) is a Nash equilibrium if for all \( x \)

\[
\langle \ell(x^*), x - x^* \rangle = \sum_k \langle \ell_{P_k}(x^*), x_{P_k} - x^*_{P_k} \rangle \geq 0
\]

I.e., for each population, every path in the support of \( x^*_{P_k} \) has minimal loss.

Rosenthal potential \( f \)

\[
f(x) = \sum_{e \in E} \int_0^{\phi_e} c_e(u) du, \phi = Mx
\]

\[
\nabla f(x) = \ell(x)
\]

\[
\mathcal{N} = \arg \min_{x \in \Delta P_1 \times \ldots \times \Delta P_K} f(x)
\]

\[
x^{(t)} \to \mathcal{N} \quad \Leftrightarrow \quad f(x^{(t)}) - f^* \to 0
\]
Average regret of population $k$

$$R_k^{(t)}(y_{\mathcal{P}_k}) = \frac{1}{t} \sum_{\tau=1}^{t} \left\langle \ell_{\mathcal{P}_k}(x^{(\tau)}), x_{\mathcal{P}_k}^{(\tau)} - y_{\mathcal{P}_k} \right\rangle$$

**Convergence of no-regret dynamics [3]**

If every population has vanishing average regret, then $\bar{x}^{(t)} = \frac{1}{t} \sum_{\tau=1}^{t} x^{(\tau)} \to \mathcal{N}$. 

**Convergence of multiplicative weights [7]**

Under multiplicative weights learning with $\eta_t \downarrow 0$, $x^{(t)} \to \mathcal{N}$. 

---


Our Results

Generalize the model:

- Observations are stochastic, losses are non Lipschitz.
- Learning is heterogeneous.
Our Results

Generalize the model:
- Observations are stochastic, losses are non Lipschitz.
- Learning is heterogeneous.

More precisely,
- Observe $\hat{\ell}(t)$, such that $\mathbb{E} \left[ \hat{\ell}(t) | \mathcal{F}_{t-1} \right] = \ell(x(t)) \text{ a.s.}$, and $\mathbb{E} \left[ \| \hat{\ell}(t) \|_2^2 \right] \leq G^2$ uniformly.
- Observation noise, or learning model with bandit feedback (form an unbiased estimator of the loss vector).
- Populations can apply different learning algorithms, in particular, different learning rates $\eta_t^k = \theta_k t^{-\alpha_k}$. 
Our Results

Generalize the model:

- Observations are stochastic, losses are non Lipschitz.
- Learning is heterogeneous.

More precisely,

- Observe $\hat{\ell}(t)$, such that $\mathbb{E} \left[ \hat{\ell}(t) \middle| \mathcal{F}_{t-1} \right] = \ell(x(t))$ a.s., and $\mathbb{E} \left[ \| \hat{\ell}(t) \|^2 \right] \leq G^2$ uniformly.
- Observation noise, or learning model with bandit feedback (form an unbiased estimator of the loss vector).
- Populations can apply different learning algorithms, in particular, different learning rates $\eta^k_t = \theta_k t^{-\alpha_k}$.

Convergence of Distributed Stochastic Mirror Descent

For $\eta^k_t = \frac{\theta_k}{t^{\alpha_k}}$, $\alpha_k \in (0, 1)$,

$$\mathbb{E} \left[ f(x(t)) \right] - f^* = \mathcal{O} \left( \sum_k \frac{\log t}{t^{\min(\alpha_k, 1-\alpha_k)}} \right)$$

In the strongly convex, homogeneous case,

$$\mathbb{E} \left[ D_\psi(x^*, x(t)) \right] = \mathcal{O} \left( t^{-\alpha} \right)$$
Stochastic Mirror Descent

\[
\text{minimize } f(x) \quad \text{convex function}
\]

\[
\text{subject to } x \in \mathcal{X} \subset \mathbb{R}^d \quad \text{convex, compact set}
\]
Stochastic Mirror Descent

\[
\text{minimize} \quad f(x) \quad \text{convex function}
\]
\[
\text{subject to} \quad x \in \mathcal{X} \subset \mathbb{R}^d \quad \text{convex, compact set}
\]

**Algorithm 2** MD Method with learning rates \((\eta_t)\)

1: \textbf{for} \(t \in \mathbb{N}\) \textbf{do}
2: \(\ell(t) \in \partial f(x(t))\)
3: \(x(t+1) = \arg\min_{x \in \mathcal{X}} \left\langle \ell(t), x \right\rangle + \frac{1}{\eta_t} D_\psi(x, x(t))\)
4: \textbf{end for}

- \(\eta_t\): learning rate
- \(D_\psi\): Bregman divergence generated by a strongly convex function \(\psi\)


Stochastic Mirror Descent

minimize \( f(x) \) \hspace{1cm} \text{convex function}
subject to \( x \in \mathcal{X} \subset \mathbb{R}^d \) \hspace{1cm} \text{convex, compact set}

Algorithm 2 MD Method with learning rates (\( \eta_t \))

1: for \( t \in \mathbb{N} \) do
2: \hspace{1cm} observe \( \ell_{P_k}^{(t)} \in \partial P_k f(x^{(t)}) \)
3: \hspace{1cm} \( x_{P_k}^{(t+1)} = \arg \min_{x \in \mathcal{X}_{P_k}} \langle \ell_{P_k}^{(t)}, x \rangle + \frac{1}{\eta_t} D_{\psi_k}(x, x_{P_k}^{(t)}) \)
4: end for

- \( \eta_t \): learning rate
- \( D_{\psi} \): Bregman divergence generated by a strongly convex function \( \psi \)


Stochastic Mirror Descent

minimize $f(x)$ convex function
subject to $x \in X \subset \mathbb{R}^d$ convex, compact set

**Algorithm 2 SMD Method with learning rates ($\eta_t$)**

1: for $t \in \mathbb{N}$ do
2: \hspace{1em} observe $\mathcal{P}_k$ with $\mathbb{E}\left[\hat{\ell}^{(t)}_{\mathcal{P}_k} | \mathcal{F}_{t-1}\right] \in \partial_{\mathcal{P}_k} f(x^{(t)})$
3: \hspace{1em} $x^{(t+1)}_{\mathcal{P}_k} = \arg \min_{x \in X_{\mathcal{P}_k}} \langle \hat{\ell}^{(t)}_{\mathcal{P}_k}, x \rangle + \frac{1}{\eta_t} D_{\psi_k}(x, x^{(t)}_{\mathcal{P}_k})$
4: end for

- $\eta_t$: learning rate
- $D_{\psi}$: Bregman divergence generated by a strongly convex function $\psi$


Bregman Divergence

Strongly convex function $\psi$

$$D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla \psi(y), x - y \rangle$$
Bregman Divergence

<table>
<thead>
<tr>
<th>Bregman Divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly convex function $\psi$</td>
</tr>
<tr>
<td>$D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla \psi(y), x - y \rangle$</td>
</tr>
</tbody>
</table>

- $\psi(x) = \frac{1}{2} \| x \|_2^2$, $D_\psi(x, y) = \frac{1}{2} \| x - y \|_2^2$ (SGD)
Bregman Divergence

Strongly convex function $\psi$

$$D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla \psi(y), x - y \rangle$$

- $\psi(x) = \frac{1}{2} \|x\|^2_2$, $D_\psi(x, y) = \frac{1}{2} \|x - y\|^2_2$ (SGD)
- $\psi(x) = -H(x) = \sum_{i=1}^d x_i \ln x_i$, $D_\psi(x, y) = D_{KL}(x, y) = \sum_{i=1}^d x_i \ln \frac{x_i}{y_i}$.

Figure: KL divergence
Example: the Hedge algorithm

\[ x_{P_k}^{(t+1)} = \arg \min_{x \in X_k} \left\langle \ell_{P_k}^{(t)}, x \right\rangle + \frac{1}{\eta_t^k} D_{KL}(x, x_{P_k}^{(t)}) \].

Hedge algorithm

Update the distribution according to observed loss

\[ x_{P}^{(t+1)} \propto x_{P}^{(t)} e^{-\eta_t^k \ell_{P}^{(t)}} \]

---


Example: the Hedge algorithm

\[ x_{k}^{(t+1)} = \arg \min_{x \in \mathcal{X}_k} \left\langle \ell(t) P_k, x \right\rangle + \frac{1}{\eta_t^k} D_{KL}(x, x_{k}^{(t)}) \].

**Hedge algorithm**

Update the distribution according to observed loss

\[ x_{p}^{(t+1)} \propto x_{p}^{(t)} e^{-\eta_t^k \ell_p^{(t)}} \]

Also known as

- Exponentially weighted average forecaster [5].

---


Example: the Hedge algorithm

\[ x_{P_k}^{(t+1)} = \arg \min_{x \in \mathcal{X}_k} \left\langle \ell_{P_k}^{(t)}, x \right\rangle + \frac{1}{\eta_t^k} D_{KL}(x, x_{P_k}^{(t)}). \]

Hedge algorithm

Update the distribution according to observed loss

\[ x_{p}^{(t+1)} \propto x_{p}^{(t)} e^{-\eta_t^k \ell_{p}^{(t)}} \]

Also known as

- Exponentially weighted average forecaster [5].
- Multiplicative weight updates [1].


Example: the Hedge algorithm

\[
x^{(t+1)}_{\mathcal{P}_k} = \arg \min_{x \in \mathcal{X}_k} \left\langle \ell^{(t)}_{\mathcal{P}_k}, x \right\rangle + \frac{1}{\eta_t^k} D_{KL}(x, x^{(t)}_{\mathcal{P}_k}).
\]

Hedge algorithm

Update the distribution according to observed loss

\[
x^{(t+1)}_p \propto x^{(t)}_p e^{-\eta_t^k \ell^{(t)}_p}
\]

Also known as
- Exponentially weighted average forecaster [5].
- Multiplicative weight updates [1].
- Exponentiated gradient descent [6].

References


Example: the Hedge algorithm

\[ x^{(t+1)}_{P_k} = \arg \min_{x \in \mathcal{X}_k} \left\langle \ell^{(t)}_{P_k}, x \right\rangle + \frac{1}{\eta^k_t} D_{KL}(x, x^{(t)}_{P_k}) . \]

Hedge algorithm

Update the distribution according to observed loss

\[ x^{(t+1)}_p \propto x^{(t)}_p e^{-\eta^k_t \ell^{(t)}_p} \]

Also known as

- Exponentially weighted average forecaster [5].
- Multiplicative weight updates [1].
- Exponentiated gradient descent [6].
- Entropic descent [2].

References


Example: the Hedge algorithm

\[ x_{p_k}^{(t+1)} = \arg \min_{x \in X_k} \langle \ell^{(t)}_{p_k}, x \rangle + \frac{1}{\eta^k_t} D_{KL}(x, x_{p_k}^{(t)}). \]

**Hedge algorithm**

Update the distribution according to observed loss

\[ x_{p}^{(t+1)} \propto x_{p}^{(t)} e^{-\eta^k_t \ell^{(t)}_{p}} \]

Also known as

- Exponentially weighted average forecaster [5].
- Multiplicative weight updates [1].
- Exponentiated gradient descent [6].
- Entropic descent [2].
- Log-linear learning


A regret bound:

\[ \sum_{\tau=t_1}^{t_2} \mathbb{E} \left[ \langle \ell_m^{(\tau)}, x_m^{(\tau)} - x_m \rangle \right] \leq \mathbb{E} \left[ D_{\psi_m}(x_m, x_m^{(t_1)}) \right] \frac{G^2}{2\mu_m} \sum_{\tau=t_1}^{t_2} \eta_m^\tau + D_m \left( \frac{1}{\eta_{t_2}^m} - \frac{1}{\eta_{t_1}^m} \right) \]


Main tool

A regret bound:

\[
\sum_{\tau=t_1}^{t_2} \mathbb{E} \left[ \langle \ell_m^{(\tau)}, x_m^{(\tau)} - x_m \rangle \right] \leq \mathbb{E} \left[ D_{\psi_m}(x_m, x_{m}^{(t_1)}) \right] + D_m \left( \frac{1}{\eta_{t_1}} - \frac{1}{\eta_{t_2}} \right) + \frac{G^2}{2\mu_m} \sum_{\tau=t_1}^{t_2} \eta_{\tau}^m
\]

From here,

- Can easily show \( \mathbb{E} \left[ f(\tilde{x}^{(t)}) \right] \to f^* \), where \( \tilde{x}^{(t)} = \frac{1}{t} \sum_{\tau=1}^{t} x^{(\tau)} \).


Main tool

A regret bound:

\[
\sum_{\tau = t_1}^{t_2} \mathbb{E} \left[ \left\langle \ell_m^{(\tau)}, x_m^{(\tau)} - x_m \right\rangle \right] \leq \frac{\mathbb{E} \left[ D_{\psi_m}(x_m, x_m^{(t_1)}) \right]}{\eta_{t_1}^m} + D_m \left( \frac{1}{\eta_{t_2}^m} - \frac{1}{\eta_{t_1}^m} \right) + \frac{G^2}{2\mu_m} \sum_{\tau = t_1}^{t_2} \eta_{\tau}^m
\]

From here,

- Can easily show \( \mathbb{E} \left[ f(\bar{x}(t)) \right] \to f^* \), where \( \bar{x}(t) = \frac{1}{t} \sum_{\tau=1}^{t} x^{(\tau)} \).
- Can show a.s. convergence \( x^{(t)} \to x^* \) if \( \sum \eta_t = \infty \) and \( \sum \eta_t^2 < \infty \)

\[
\mathbb{E} \left[ D_{\psi}(x^*, x^{(\tau+1)}) | \mathcal{F}_{\tau-1} \right] \leq D_{\psi}(x^*, x^{(\tau)}) - \eta_{\tau} (f(x^{(\tau)}) - f^*) + \frac{\eta_{\tau}^2}{2\mu} \mathbb{E} \left[ \| \hat{\ell}^{(\tau)} \|^2_* | \mathcal{F}_{\tau-1} \right]
\]

---


Main tool

A regret bound:

\[
\sum_{\tau=t_1}^{t_2} \mathbb{E} \left[ \langle \ell_m^{(\tau)}, x_m^{(\tau)} - x_m \rangle \right] \leq \mathbb{E} \left[ D_{\psi_m}(x_m, x_m^{(t_1)}) \right] + D_m \left( \frac{1}{\eta_{t_1}} - \frac{1}{\eta_{t_2}} \right) + \frac{G^2}{2\mu_m} \sum_{\tau=t_1}^{t_2} \eta_m^2
\]

From here,

- Can easily show \( \mathbb{E} \left[ f(\bar{x}(t)) \right] \to f^* \), where \( \bar{x}(t) = \frac{1}{t} \sum_{\tau=1}^{t} x^{(\tau)} \).
- Can show a.s. convergence \( x^{(t)} \to \mathcal{X}^* \) if \( \sum \eta_t = \infty \) and \( \sum \eta_t^2 < \infty \)

\[
\mathbb{E} \left[ D_{\psi}(\mathcal{X}^*, x^{(\tau+1)}) | \mathcal{F}_{\tau-1} \right] \leq D_{\psi}(\mathcal{X}^*, x^{(\tau)}) - \eta_\tau (f(x^{(\tau)}) - f^*) + \frac{\eta_\tau^2}{2\mu} \mathbb{E} \left[ \|\hat{\ell}^{(\tau)}\|^2_2 | \mathcal{F}_{\tau-1} \right]
\]

\( D_{\psi}(\mathcal{X}^*, x^{(\tau)}) \) is an almost super martingale \([10]\), so \( D_{\psi}(\mathcal{X}^*, x^{(\tau)}) \) converges a.s. and \( \sum_{\tau} \eta_\tau (f(x^{(\tau)}) - f^*) < \infty \) a.s.

Generalizes a known result in stochastic approximation, e.g. \([4]\) (for SGD, for strictly convex functions).

\[10\]H. Robbins and D. Siegmund. *A convergence theorem for non negative almost supermartingales and some applications.*  
*Optimizing Methods in Statistics*, 1971

\[4\]Léon Bottou. *Online algorithms and stochastic approximations.*  
Main tools and results

- To show convergence $\mathbb{E} \left[ f(x^{(t)}) \right] \to f^*$, generalize the technique of Shamir et al. [11] (for SGD, $\alpha = \frac{1}{2}$).

Convergence of Distributed Stochastic Mirror Descent

For $\eta_t^k = \frac{\theta_k}{t^{\alpha_k}}$, $\alpha_k \in (0, 1)$,

$$\mathbb{E} \left[ f(x^{(t)}) \right] - f^* = O \left( \sum \log \frac{t}{t^{\min(\alpha_k, 1-\alpha_k)}} \right)$$

Non-smooth, non-strongly convex.

Example: routing game with non strongly convex potential

Figure: A non strongly convex example.
Learning model: (smoothed) entropic mirror descent, with $\eta_t^k = \theta_k t^{-\alpha_k}$
Example: routing game with non strongly convex potential

\[
\frac{\theta_k}{t^{\alpha_k}}, \quad \alpha_k \in (0, 1), \quad \mathbb{E} \left[ f(x(t)) \right] - f^* = O \left( \sum_k \frac{\log t}{t^{\min(\alpha_k, 1 - \alpha_k)}} \right)
\]
Example: routing game with non strongly convex potential

\[ E \left[ f(x(\tau)) \right] - f^* = O \left( \sum_k \frac{\log t}{t^{\min(\alpha_k, 1-\alpha_k)}} \right) \]

**Figure:** Potential values.
Example: routing game with non strongly convex potential

Figure: Potential values.

For $\frac{\theta_k}{t^{\alpha_k}}$, $\alpha_k \in (0, 1)$, $E[f(x(t))] - f^* = O\left(\sum_k \frac{\log t}{t^{\min(\alpha_k, 1-\alpha_k)}}\right)$
Example: strongly convex potential

Figure: A strongly convex example.
Learning model: (smoothed) entropic mirror descent, with $\eta_t = t^{-1}$
Example: routing game with non strongly convex potential

![Graph showing potential values.]

**Figure:** Potential values.
\[
\mathbb{E} [D_{\psi}(x^*, x(t))] = O(t^{-1})
\]
Conclusion

Summary

- A more realistic model: stochastic observations, non-Lipschitz, heterogeneous learning.
- Convergence bounds for Stochastic Mirror Descent, with heterogeneous learning rates.
- Convergence of $x^{(t)}$ instead of $\bar{x}^{(t)}$. 
Conclusion

Summary

- A more realistic model: stochastic observations, non-Lipschitz, heterogeneous learning.
- Convergence bounds for Stochastic Mirror Descent, with heterogeneous learning rates.
- Convergence of $x(t)$ instead of $\bar{x}(t)$.

Current and future work

- Model of learning at the player level.
- Estimation of model parameters (e.g. learning rate)
- Optimal control on top of this behavioral model
Thank you.

eecs.berkeley.edu/~walid
References I


References II


