Convergence of Heterogeneous Distributed Learning In Stochastic Routing Game

Syrine Krichene Walid Krichene Roy Dong Alexandre Bayen

September 30, 2015

Outline

Routing game

Used to model congestion in

- Transportation networks
- Communication networks

Routing game

Used to model congestion in

- Transportation networks
- Communication networks

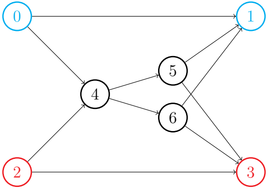


Figure: Example network

- Directed graph (V, E)
- Population k: paths $\mathcal{P}_k$

Routing game

Used to model congestion in

- Transportation networks
- Communication networks

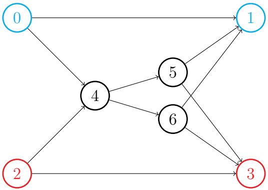


Figure: Example network

- 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

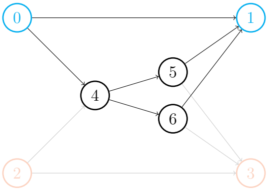


Figure: Example network

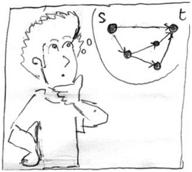
- Directed graph (V, E)
- Population k: paths $\mathcal{P}_k$
- Population distribution over paths $x_{\mathcal{P}_k} \in \Delta^{\mathcal{P}_k}$
- Loss on path p: $\ell_p(x) = \sum_{e \in p} c_e(\phi_e)$

Online learning model

Online Learning Model

```
for t ∈ ℕ do
    Play p ~ x_{P_k}^{(t)}
    Discover ℓ_{P_k}^{(t)}
    Update x_{P_k}^{(t+1)}
end for
```

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

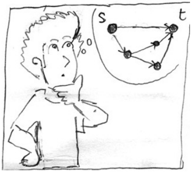


Online learning model

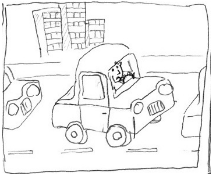
Online Learning Model

for $t \in \mathbb{N}$ **do**
    Play $p \sim x_{\mathcal{P}_k}^{(t)}$
    Discover $\ell_{\mathcal{P}_k}^{(t)}$
    Update $x_{\mathcal{P}_k}^{(t+1)}$
end for

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



Sample $p \sim x_{\mathcal{P}_1}^{(t)}$



Online learning model

Online Learning Model

for $t \in \mathbb{N}$ **do**
  Play $p \sim x_{\mathcal{P}_k}^{(t)}$
  Discover $\ell_{\mathcal{P}_k}^{(t)}$
  Update $x_{\mathcal{P}_k}^{(t+1)}$
end for

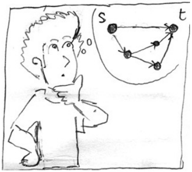
$x_{\mathcal{P}_1}^{(t)} \in \Delta^{\mathcal{P}_1}$ Sample $p \sim x_{\mathcal{P}_1}^{(t)}$ Discover ${\ell_{\mathcal{P}_1}}^{(t)}$

Online learning model

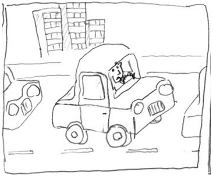
Online Learning Model

for $t \in \mathbb{N}$ **do**
  Play $p \sim x_{\mathcal{P}_k}^{(t)}$
  Discover $\ell_{\mathcal{P}_k}^{(t)}$
  Update $x_{\mathcal{P}_k}^{(t+1)}$
end for

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

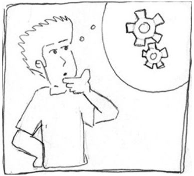


Sample $p \sim x_{\mathcal{P}_1}^{(t)}$



Discover $\ell_{\mathcal{P}_1}{}^{(t)}$

Update $x_{\mathcal{P}_1}^{(t+1)}$



Convergence to Nash equilibria

Nash equilibrium

$x^\star$ is a Nash equilibrium if for all x

$$\langle \ell(x^\star), x - x^\star \rangle = \sum_k \left\langle \ell_{\mathcal{P}_k}(x^\star), x_{\mathcal{P}_k} - x^\star_{\mathcal{P}_k} \right\rangle \geq 0$$

I.e., for each population, every path in the support of $x^\star_{\mathcal{P}_k}$ has minimal loss.

Convergence to Nash equilibria

Nash equilibrium

$x^\star$ is a Nash equilibrium if for all x

$$\langle \ell(x^\star), x - x^\star \rangle = \sum_k \langle \ell_{\mathcal{P}_k}(x^\star), x_{\mathcal{P}_k} - x^\star_{\mathcal{P}_k} \rangle \geq 0$$

I.e., for each population, every path in the support of $x^\star_{\mathcal{P}_k}$ has minimal loss.

Rosenthal potential f

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

$$\nabla f(x) = \ell(x)$$

$$\mathcal{N} = \arg\min_{x \in \Delta^{\mathcal{P}_1} \times \cdots \times \Delta^{\mathcal{P}_K}} f(x)$$

$$x^{(t)} \to \mathcal{N} \qquad \Leftrightarrow \qquad f(x^{(t)}) - f^\star \to 0$$

Previous Results

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 $\eta_t \downarrow 0$, $x^{(t)} \to \mathcal{N}$.

---

[3]Avrim Blum, Eyal Even-Dar, and Katrina Ligett. Routing without regret: on convergence to nash equilibria of regret-minimizing algorithms in routing games. In *Proceedings of the twenty-fifth annual ACM symposium on Principles of distributed computing*, PODC '06, pages 45–52, New York, NY, USA, 2006. ACM

[7]Robert Kleinberg, Georgios Piliouras, and Eva Tardos. Multiplicative updates outperform generic no-regret learning in congestion games. In *Proceedings of the 41st annual ACM symposium on Theory of computing*, pages 533–542. ACM, 2009

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 $\mathbb{E}\left[\hat{\ell}^{(t)}|\mathcal{F}_{t-1}\right] = \ell(x^{(t)})$ 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 $\eta_t^k = \theta_k t^{-\alpha_k}$.

Our Results

Generalize the model:

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

More precisely,

- Observe $\hat{\ell}^{(t)}$, such that $\mathbb{E}\left[\hat{\ell}^{(t)} | \mathcal{F}_{t-1}\right] = \ell(x^{(t)})$ 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 $\eta_t^k = \theta_k t^{-\alpha_k}$.

Convergence of Distributed Stochastic Mirror Descent

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

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

In the strongly convex, homogeneous case,

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

Stochastic Mirror Descent

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

[9]A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization.*
Wiley-Interscience series in discrete mathematics. Wiley, 1983

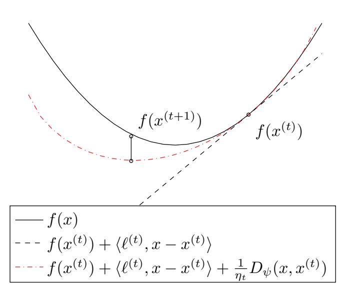
[8]A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming.
SIAM Journal on Optimization, 19(4):1574–1609, 2009

Stochastic Mirror Descent

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

Algorithm 2 MD Method with learning rates (η_t)

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



- η_t: learning rate
- D_ψ: Bregman divergence generated by a strongly convex function ψ

[9]A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization.* Wiley-Interscience series in discrete mathematics. Wiley, 1983

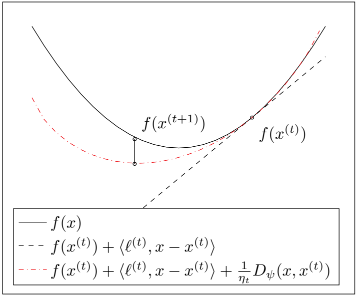
[8]A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009

Stochastic Mirror Descent

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

Algorithm 2 MD Method with learning rates (η_t)

1: **for** $t \in \mathbb{N}$ **do**
2: observe $\ell^{(t)}_{\mathcal{P}_k} \in \partial_{\mathcal{P}_k} f(x^{(t)})$
3: $x^{(t+1)}_{\mathcal{P}_k} = \arg\min_{x \in \mathcal{X}_{\mathcal{P}_k}} \left\langle \ell^{(t)}_{\mathcal{P}_k}, x \right\rangle + \frac{1}{\eta^k_t} D_{\psi_k}(x, x^{(t)}_{\mathcal{P}_k})$
4: **end for**



- η_t: learning rate
- D_ψ: Bregman divergence generated by a strongly convex function ψ

[9]A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization.*
Wiley-Interscience series in discrete mathematics. Wiley, 1983

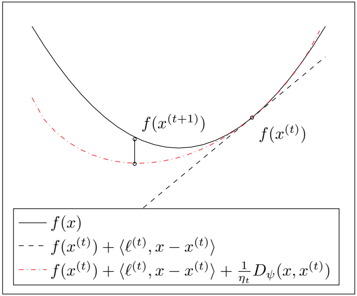
[8]A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming.
SIAM Journal on Optimization, 19(4):1574–1609, 2009

Stochastic Mirror Descent

$$\begin{array}{lll} \text{minimize} & f(x) & \text{convex function} \\ \text{subject to} & x \in \mathcal{X} \subset \mathbb{R}^d & \text{convex, compact set} \end{array}$$

Algorithm 2 SMD Method with learning rates (η_t)

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



- η_t: learning rate
- D_ψ: Bregman divergence generated by a strongly convex function ψ

[9]A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization.*
Wiley-Interscience series in discrete mathematics. Wiley, 1983

[8]A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming.
SIAM Journal on Optimization, 19(4):1574–1609, 2009

Bregman Divergence

Bregman Divergence

Strongly convex function ψ

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

Bregman Divergence

Bregman Divergence

Strongly convex function ψ

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

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

Bregman Divergence

Bregman Divergence

Strongly convex function ψ

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

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

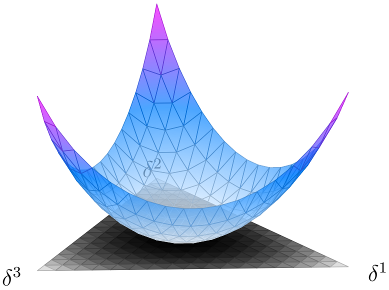


Figure: KL divergence

Example: the Hedge algorithm

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

Hedge algorithm

Update the distribution according to observed loss

$$x_p^{(t+1)} \propto x_p^{(t)} e^{-\eta_t^k \ell_p^{(t)}}$$

[5]Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games.* Cambridge University Press, 2006

[1]Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012

[6]Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1 – 63, 1997

[2]Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient

Example: the Hedge algorithm

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

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].

---

[5]Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games.* Cambridge University Press, 2006

[1]Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012

[6]Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1 – 63, 1997

[2]Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient

Example: the Hedge algorithm

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

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].

---

[5]Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games.* Cambridge University Press, 2006

[1]Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012

[6]Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1 – 63, 1997

[2]Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient

Example: the Hedge algorithm

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

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].
- Exponentiated gradient descent [6].

---

[5]Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games.* Cambridge University Press, 2006

[1]Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012

[6]Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1 – 63, 1997

[2]Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient

Example: the Hedge algorithm

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

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].
- Exponentiated gradient descent [6].
- Entropic descent [2].

---

[5]Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games.*
Cambridge University Press, 2006

[1]Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications.
Theory of Computing, 8(1):121–164, 2012

[6]Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors.
Information and Computation, 132(1):1 – 63, 1997

[2]Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient

Example: the Hedge algorithm

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

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].
- Exponentiated gradient descent [6].
- Entropic descent [2].
- Log-linear learning

[5]Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games.* Cambridge University Press, 2006

[1]Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012

[6]Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1 – 63, 1997

[2]Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient

Main tool

A regret bound:

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

---

[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.
In David Saad, editor, *Online Learning and Neural Networks*. Cambridge University Press, Cambridge, UK, 1998

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 $\mathbb{E}\left[f(\bar{x}^{(t)})\right] \to f^\star$, where $\bar{x}^{(t)} = \frac{1}{t}\sum_{\tau=1}^{t} x^{(\tau)}$.

[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.
In David Saad, editor, *Online Learning and Neural Networks*. Cambridge University Press, Cambridge, UK, 1998

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

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

[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.
In David Saad, editor, *Online Learning and Neural Networks*. Cambridge University Press, Cambridge, UK, 1998.

Main tool

A regret bound:

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

From here,

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

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

$D_\psi(\mathcal{X}^\star, x^{(\tau)})$ is an almost super martingale [10], so $D_\psi(\mathcal{X}^\star, x^{(\tau)})$ converges a.s. and $\sum_\tau \eta_\tau(f(x^{(\tau)}) - f^\star) < \infty$ 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.
In David Saad, editor, *Online Learning and Neural Networks*. Cambridge University Press, Cambridge, UK, 1998.

Main tools and results

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

Convergence of Distributed Stochastic Mirror Descent

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

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

Non-smooth, non-strongly convex.

---

[11]Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *ICML*, pages 71–79, 2013

Example: routing game with non strongly convex potential

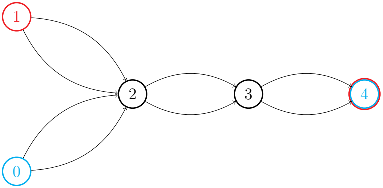


Figure: A non strongly convex example.
Learning model: (smoothed) entropic mirror descent, with $\eta_t^k = \theta_k t^{-\alpha_k}$

Example: routing game with non strongly convex potential

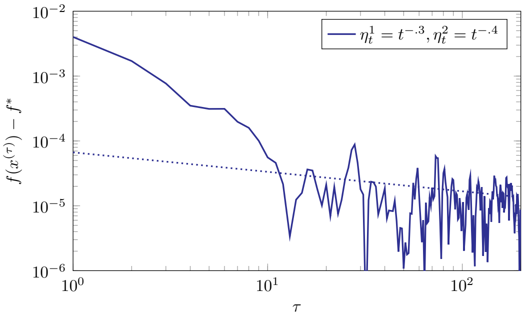


Figure: Potential values.

For $\frac{\theta_k}{t^{\alpha_k}},\ \alpha_k \in (0,1)$, $\mathbb{E}\left[f(x^{(t)})\right] - f^\star = O\left(\sum_k \frac{\log t}{t^{\min(\alpha_k, 1-\alpha_k)}}\right)$

Example: routing game with non strongly convex potential

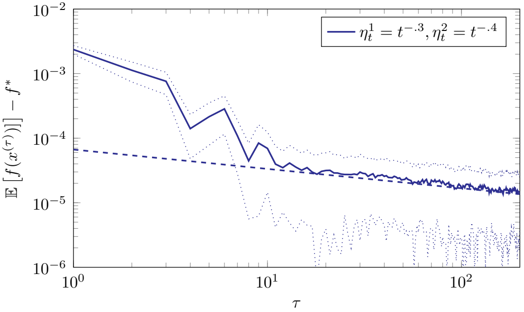


Figure: Potential values.

For $\frac{\theta_k}{t^{\alpha_k}}$, $\alpha_k \in (0,1)$, $\mathbb{E}\left[f(x^{(t)})\right] - f^\star = O\left(\sum_k \frac{\log t}{t^{\min(\alpha_k, 1-\alpha_k)}}\right)$

Example: routing game with non strongly convex potential

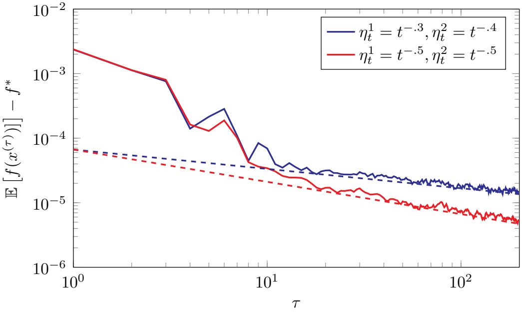


Figure: Potential values.

For $\frac{\theta_k}{t^{\alpha_k}}$, $\alpha_k \in (0,1)$, $\mathbb{E}\left[f(x^{(t)})\right] - f^\star = O\left(\sum_k \frac{\log t}{t^{\min(\alpha_k, 1-\alpha_k)}}\right)$

Example: strongly convex potential

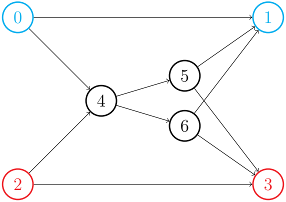


Figure: A strongly convex example.
Learning model: (smoothed) entropic mirror descent, with $\eta_t = t^{-1}$

Example: routing game with non strongly convex potential

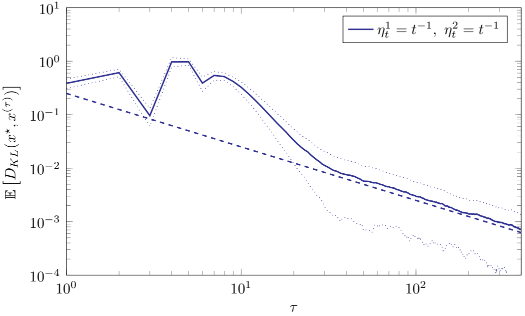


Figure: Potential values.
$\mathbb{E}\left[D_{\psi}(x^{\star}, x^{(t)})\right] = O\left(t^{-1}\right)$

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

[1] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

[2] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, 31(3): 167–175, May 2003.

[3] Avrim Blum, Eyal Even-Dar, and Katrina Ligett. Routing without regret: on convergence to nash equilibria of regret-minimizing algorithms in routing games. In *Proceedings of the twenty-fifth annual ACM symposium on Principles of distributed computing*, PODC '06, pages 45–52, New York, NY, USA, 2006. ACM.

[4] Léon Bottou. Online algorithms and stochastic approximations. In David Saad, editor, *Online Learning and Neural Networks*. Cambridge University Press, Cambridge, UK, 1998. revised, oct 2012.

[5] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

References II

[6] Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132 (1):1 – 63, 1997.

[7] Robert Kleinberg, Georgios Piliouras, and Eva Tardos. Multiplicative updates outperform generic no-regret learning in congestion games. In *Proceedings of the 41st annual ACM symposium on Theory of computing*, pages 533–542. ACM, 2009.

[8] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[9] A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience series in discrete mathematics. Wiley, 1983.

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

[11] Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *ICML*, pages 71–79, 2013.