Stochastic optimization with non-i.i.d. noise

Alekh Agarwal  John C. Duchi
UC Berkeley
Basic setup

- Want to minimize over $x \in \mathcal{X}$
  \[
  f(x) := \mathbb{E}_\Pi[F(x; \omega)] = \int_\Omega F(x; \omega) d\Pi(\omega)
  \]

- Cannot compute $f$ directly
Basic setup

- Want to minimize over $x \in \mathcal{X}$

$$f(x) := \mathbb{E}_\Pi[F(x; \omega)] = \int_\Omega F(x; \omega) d\Pi(\omega)$$

- Cannot compute $f$ directly
- Classical Stochastic Approximations requires i.i.d. samples from $\Pi$
Basic setup

- Want to minimize over $x \in \mathcal{X}$

$$f(x) := \mathbb{E}_\Pi[F(x; \omega)] = \int_\Omega F(x; \omega) d\Pi(\omega)$$

- Cannot compute $f$ directly
- Classical Stochastic Approximations requires i.i.d. samples from $\Pi$

$\omega_1$ and $\omega_2$ should be i.i.d
Basic setup

- Want to minimize over $x \in \mathcal{X}$

$$f(x) := \mathbb{E}_\Pi[F(x; \omega)] = \int_\Omega F(x; \omega) d\Pi(\omega)$$

- Cannot compute $f$ directly
- Classical Stochastic Approximations requires i.i.d. samples from $\Pi$
- What if $\Pi$ cannot be sampled from?
- What if samples are not i.i.d.?
Observe sequence of functions $F(\cdot; \omega_1), \ldots, F(\cdot; \omega_T)$

No statistical structure assumed on $\omega_t$

Low regret to best fixed point

$$\sum_{t=1}^{T} F(x_t; \omega_t) - \min_{x \in \mathcal{X}} \sum_{t=1}^{T} F(x; \omega_t).$$
Online Convex Optimization

- Observe sequence of functions $F(\cdot; \omega_1), \ldots, F(\cdot; \omega_T)$
- No statistical structure assumed on $\omega_t$
- Low regret to best fixed point

$$\sum_{t=1}^{T} F(x_t; \omega_t) - \min_{x \in \mathcal{X}} \sum_{t=1}^{T} F(x; \omega_t).$$

- Implies low optimization error on $f$ when $\omega_t$ are i.i.d. from $\Pi$ [Cesa-Bianchi et al. '04]
- What if $\Pi$ cannot be sampled from?
- What if samples are not i.i.d.?
I.I.D. sampling not always possible

Non-i.i.d. noise arises in several contexts

- Time series data
- Financial data
- Robotics and reinforcement learning
- Control and dynamical systems
- Optimization over combinatorial spaces
Formal setup

- Stochastic process $\mathbb{P}^t$
- Noise sequence $(\omega_1, \ldots, \omega_T) \sim \mathbb{P}$
- Stationary distribution $\Pi$ such that $\mathbb{P}^t \rightarrow \Pi$
- $\mathbb{P}^t = \Pi$ for all $t$ in i.i.d. scenario
Formal setup

- Stochastic process $\mathbb{P}^t$
- Noise sequence $(\omega_1, \ldots, \omega_T) \sim \mathbb{P}$
- Stationary distribution $\Pi$ such that $\mathbb{P}^t \to \Pi$
- $\mathbb{P}^t = \Pi$ for all $t$ in i.i.d. scenario

Definition ($\phi$-mixing)

Let $P^t_s$ be the distribution at time $t$, conditioning on first $s$ samples. The $\phi$-mixing coefficient is defined as

$$\phi(k) = \sup_{t \in \mathbb{N}} \left\{ d_{TV}(P^t_{t+k}, \Pi) \right\}. $$

- $\phi(1) = 0$ for i.i.d.
Example 1: Markov Incremental Gradient Descent (MIGD) [Johansson et al., ’09]

- Node $i$ has local function $F(x; i) = f_i(x)$
- Want to minimize $f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x)$
- Token $i_t$ moves on graph according to random walk
Example 1: Markov Incremental Gradient Descent (MIGD) [Johansson et al., '09]

- Node $i$ has local function $F(x; i) = f_i(x)$
- Want to minimize $f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x)$
- Token $i_t$ moves on graph according to random walk
Example 1: Markov Incremental Gradient Descent (MIGD) [Johansson et al., '09]

- Node $i$ has local function $F(x; i) = f_i(x)$
- Want to minimize $f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x)$
- Token $i_t$ moves on graph according to random walk
Example 1: Markov Incremental Gradient Descent (MIGD) [Johansson et al., ’09]

- Node $i$ has local function $F(x; i) = f_i(x)$
- Want to minimize $f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x)$
- Token $i_t$ moves on graph according to random walk
Example 1: Markov Incremental Gradient Descent (MIGD) [Johansson et al., ’09]

- Node $i$ has local function $F(x; i) = f_i(x)$
- Want to minimize $f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x)$
- Token $i_t$ moves on graph according to random walk
- $\mathbb{P}^{t} \to \frac{I}{m}$, $\phi(k) \leq (1 - \sigma_2(P))^k$
- $\sigma_2(P)$ is the second largest singular value of transition matrix
Example 1: Markov Incremental Gradient Descent (MIGD) [Johansson et al., ’09]

- Node $i$ has local function $F(x; i) = f_i(x)$
- Want to minimize $f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x)$
- Token $i_t$ moves on graph according to random walk
  - $P^t \rightarrow \frac{1}{m}$, $\phi(k) \leq (1 - \sigma_2(P))^k$
- $f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x)$
- Observe $g_t = \nabla F(x(t); i_t) = \nabla f_{i_t}(x(t))$
- Gradient update: $x(t + 1) = \Pi_x(x(t) - \eta_t g_t)$
Example 2: Autoregressive process

- $\omega_t = A\omega_{t-1} + z_t$ where $z_t \sim \mathcal{N}(\mu, \Sigma)$
- Model for time series, control system dynamics etc.
- Observe samples $\omega_t, \nu_t$
- $F(x; (\omega, \nu)) = (\nu - x^T\omega)^2$
- $\Pi$ is $\mathcal{N}((I - A)^{-1}\mu, \Sigma_*)$
Stable online algorithms

- Assume access to low-regret online algorithm
  \[ \sum_{t=1}^{T} F(x_t; \omega_t) - F(x^*; \omega_t) \leq R_T \quad \forall x^* \in X \]

- Assume stability of iterates
  \[ \|x(t) - x(t-1)\| \leq \kappa(t) \]

- Online Gradient Descent (OGD) and Regularized Dual Averaging (RDA) satisfy both above conditions
Convergence rate for convex losses

- Assume iterates $x_1, \ldots, x_T$ from a stable online algorithm
- Define the average

$$\hat{x}_T = \frac{1}{T} \sum_{i=1}^{T} x(i)$$

**Theorem**

Let the functions $F(\cdot; \omega)$ be convex and $G$-Lipschitz and for any $x, y \in \mathcal{X}$, $\|x - y\| \leq R$. Then for any $\tau$ and $x^* \in \mathcal{X}$ with probability at least $1 - \delta$,

$$f(\hat{x}_T) - f(x^*) \leq \frac{R_T}{T} + 3GR\sqrt{\frac{\tau}{T} \log \frac{\tau}{\delta}} + \frac{(\tau - 1)G}{T} \sum_{t=1}^{T} \kappa(t) + \phi(\tau)GR$$

- Avg. Regret
- Deviation bound
- Non-i.i.d. Penalty
Convergence rate for convex losses

- Assume iterates $x_1, \ldots, x_T$ from a stable online algorithm
- Define the average
  \[
  \hat{x}_T = \frac{1}{T} \sum_{i=1}^{T} x(i)
  \]

**Theorem**

Let the functions $F(\cdot; \omega)$ be convex and $G$-Lipschitz and for any $x, y \in \mathcal{X}$, $\|x - y\| \leq R$. Then for any $\tau$ and $x^* \in \mathcal{X}$ with probability at least $1 - \delta$,

\[
 f(\hat{x}_T) - f(x^*) \leq \frac{\mathcal{R}_T}{T} + 3GR \sqrt{\frac{\tau}{T} \log \frac{\tau}{\delta}} + \frac{(\tau - 1)G}{T} \sum_{t=1}^{T} \kappa(t) + \phi(\tau)GR
\]

- Set $\tau = 1$ and $\phi(1) = 0$ to recover i.i.d. result
Specialization to particular algorithms

- Consider OGD or RDA with $\eta_t \propto 1/\sqrt{t}$
- Both OGD and RDA satisfy

$$ R_T = \mathcal{O} \left( GR\sqrt{T} \right) \quad \text{and} \quad \kappa(t) = \mathcal{O} \left( \frac{R}{\sqrt{t}} \right) $$

**Corollary**

*Let the iterates $x(1), \ldots, x(T)$ be generated according to OGD or RDA. Then for a universal constant $C$ with probability at least $1 - \delta$,*

$$ f(\hat{x}_T) - f(x^*) \leq \frac{C}{\sqrt{T}} \cdot \inf_{\tau \in \mathbb{N}} \left[ GR(\tau - 1) + GR \sqrt{\tau \log \frac{\tau}{\delta}} + \sqrt{T} \phi(\tau) GR \right] $$

Almost sure convergence as long as $\phi(k) \to 0$ when $k \to \infty$
Specialization to geometric mixing

- Several examples satisfy $\phi(k) = \phi_0 \exp(-\phi_1 k)$
- E.g.: Finite-state Markov Chains

**Corollary**

*Under geometric mixing, the iterates of OGD or RDA satisfy the following guarantee for a universal constant $C$ with probability at least $1 - \delta$*

$$f(\hat{x}_T) - f(x^*) \leq C \cdot \left[ \frac{GR}{\sqrt{T}} \frac{\log T}{\phi_1} + \frac{GR}{\sqrt{T}} \sqrt{\frac{\log T}{\phi_1} \log \frac{\log T}{\delta}} \right]$$

Almost same as i.i.d. result [Cesa-Bianchi et al., (2004)]

$$f(\hat{x}_T) - f(x^*) \leq C \cdot \left[ \frac{GR}{\sqrt{T}} \sqrt{\log \frac{1}{\delta}} \right]$$
Simulation

- Convergence rate on 25-dimensional AR process for OGD algorithm

![Graph showing relative regression error vs. number of samples](image-url)
Example: Markov Incremental Gradient Descent (MIGD) [Johansson et al., ’09]

- Node \( i \) has local function \( f_i(x) \)
- Token \( i_t \) moves on graph according to random walk
Example: Markov Incremental Gradient Descent (MIGD) [Johansson et al., ’09]

- Node $i$ has local function $f_i(x)$
- Token $i_t$ moves on graph according to random walk
Example: Markov Incremental Gradient Descent (MIGD) [Johansson et al., ’09]

- Node $i$ has local function $f_i(x)$
- Token $i_t$ moves on graph according to random walk
Example: Markov Incremental Gradient Descent (MIGD) [Johansson et al., ’09]

- Node $i$ has local function $f_i(x)$
- Token $i_t$ moves on graph according to random walk
Example: Markov Incremental Gradient Descent (MIGD) [Johansson et al., ’09]

- Node $i$ has local function $f_i(x)$
- Token $i_t$ moves on graph according to random walk
- Observe $g_t = \nabla f_{i_t}(x(t))$
- Gradient update: $x(t + 1) = \Pi x(x(t) - \eta_t g_t)$
Convergence rate of MIGD

- $\mathbb{P}^t \to \frac{1}{m}, \phi(k) \leq (1 - \sigma_2(P))^k$
- $f(x) = \frac{1}{m} \sum_{i=1}^m f_i(x)$
- Stability of OGD implies MIGD converges

**Corollary**

After $T$ steps of MIGD, with probability at least $1 - \delta$

$$\frac{1}{m} \sum_{i=1}^m f_i(\hat{x}_T) - \frac{1}{m} \sum_{i=1}^m f_i(x^*) = O \left( \frac{RG}{\sqrt{T}} \sqrt{\frac{\log \frac{1}{\delta}}{1 - \sigma_2(P)}} \right)$$
Convergence is fastest for Expanders and slowest for Cycles as predicted by theory based on spectral gaps.
A function $h$ is $\lambda$-strongly convex if

$$h(x) \geq h(y) + \nabla h(y)^T (x - y) + \frac{\lambda}{2} \|x - y\|^2, \quad \forall x, y \in \mathcal{X}$$

Examples

- $\frac{1}{2} \|x\|^2$ is $1$-strongly convex
- Regularized objectives such as SVM, ridge regression, regularized logistic, entropy-based regularization etc.
Better guarantees for strongly convex losses

- Assume $F(\cdot; \omega)$ is $\lambda$-strongly convex

- OGD and RDA obtain regret $O\left(\frac{G^2 \log T}{\lambda}\right)$

**Theorem**

Let $F(\cdot; \omega)$ be $G$-Lipschitz and $\lambda$-strongly convex. Let $x_1, \ldots, x_T$ be generated according to OGD or RDA. Then with probability at least $1 - \delta \log T$,

$$f(\hat{x}_T) - f(x^*) \leq O\left(\frac{G^2}{\lambda T} \inf_{\tau \in \mathbb{N}} \left[ \tau \left(\log T + \log \frac{T}{\delta}\right) + T \phi(\tau) \right]\right)$$

Set $\tau = 1, \phi(1) = 0$ for i.i.d.

Almost as good as i.i.d. for geometric mixing.
Better guarantees for strongly convex losses

- Assume $F(\cdot; \omega)$ is $\lambda$-strongly convex
- OGD and RDA obtain regret $O\left(\frac{G^2 \log T}{\lambda}\right)$

**Theorem**

Let $F(\cdot; \omega)$ be $G$-Lipschitz and $\lambda$-strongly convex. Let $x_1, \ldots, x_T$ be generated according to OGD or RDA. Then with probability at least $1 - \delta \log T$,

$$f(\hat{x}_T) - f(x^*) \leq O\left(\frac{G^2}{\lambda T} \inf_{\tau \in \mathbb{N}} \left[ \tau \left(\log T + \log \frac{\tau}{\delta}\right) + T \phi(\tau)\right]\right)$$

- Faster than $O(1/\sqrt{T})$ bound for convex losses
- Set $\tau = 1, \phi(1) = 0$ for i.i.d.
- Almost as good as i.i.d. for geometric mixing
Many problems of the form $F(x; \omega) = \ell(\langle x, \omega \rangle)$
- Logistic regression, boosting, least-squares regression etc.
Fast rates for linear prediction

- Many problems of the form $F(x; \omega) = \ell(\langle x, \omega \rangle)$
  - Logistic regression, boosting, least-squares regression etc.
- $F(x; \omega)$ cannot be strongly convex
- Often $\ell(u)$ is strongly convex in $u$
- We prove fast rates of $O(1/T)$ under strong convexity of $\ell$
- Requires a slight modification of existing algorithms
Conclusions

- Convergence rates for stochastic optimization with non-i.i.d. data
- Theory applies to all *stable online algorithms*
- Analysis covers large class of mixing processes
- Similar rates as i.i.d. data under geometric mixing
- Fast rates for strongly convex objectives and linear prediction
Extensions

- Results also provided for weaker $\beta$-mixing assumption
- Other applications:
  - Learning a ranking function from partially ordered samples
  - Optimization with MCMC samples from a combinatorial space
  - Optimizing the parameters of a queuing system
  - Simulation experiments with pseudo-random numbers
References