Convex Optimization

(EE227A: UC Berkeley)

Lecture 19
(Stochastic optimization)

02 Apr, 2013

Suvrit Sra

Admin

- ♠ HW3 due **4/04/2013**
- ♠ HW4 on bSpace later today–due **4/18/2013**
- ♠ Project report (4 pages) due on: **11th April**
- ♠ LaTeX template for projects on bSpace

Recap

- ♠ Convex sets, functions
- ♠ Convex models, LP, QP, SOCP, SDP
- ♠ Subdifferentials, basic optimality conditions
- ♠ Weak duality
- ♠ Lagrangians, strong duality, KKT conditions

- ♠ Subgradient method
- ♠ Gradient descent, feasible descent
- ♠ Optimal gradients methods
- ♠ Constrained problems, conditional gradient
- ♠ Nonsmooth problems, proximal methods
- ♠ Proximal splitting, Douglas-Rachford
- ♠ Monotone operators, product-space trick
- ♠ Incremental gradient methods

Incremental methods

$$\boxed{\min \quad [f(x) = \sum_i f_i(x)] + r(x)}$$

$$x^{k+1} \quad = \quad x^k - \alpha_k g^{i(k)}, \quad g^{i(k)} \in \partial f_{i(k)}(x^k)$$

Incremental methods

$$\boxed{\min \quad [f(x) = \textstyle\sum_i f_i(x)] + r(x)}$$

$$\begin{aligned} x^{k+1} &= x^k - \alpha_k g^{i(k)}, \quad g^{i(k)} \in \partial f_{i(k)}(x^k) \\ x^{k+1} &= \mathrm{prox}_{\alpha_k f_{i(k)}}(x^k), \qquad k = 0, 1, \ldots \end{aligned}$$

Incremental methods

$$\boxed{\min \quad [f(x) = \textstyle\sum_i f_i(x)] + r(x)}$$

$$\begin{aligned}
x^{k+1} &= x^k - \alpha_k g^{i(k)}, \quad g^{i(k)} \in \partial f_{i(k)}(x^k) \\
x^{k+1} &= \mathrm{prox}_{\alpha_k f_{i(k)}}(x^k), \qquad k = 0, 1, \ldots \\
x^{k+1} &= \mathrm{prox}_{\alpha_k r}\Big(x^k - \eta_k \sum_{i=1}^{m} \nabla f_i(z^i)\Big), \quad k = 0, 1, \ldots, \\
& \quad z^1 = x^k \\
& \quad z^{i+1} = z^i - \alpha_k \nabla f_i(z^i), \quad i = 1, \ldots, m-1.
\end{aligned}$$

Incremental methods

$$\boxed{x^{k+1} = P_{\mathcal{X}}(x^k - \alpha_k \nabla f_{i(k)}(x^k))}$$

Choices of $i(k)$

- *Cyclic:* $i(k) = 1 + (k \mod m)$
- *Randomized:* Pick $i(k)$ uniformly from $\{1, \ldots, m\}$

Incremental methods

$$\boxed{x^{k+1} = P_{\mathcal{X}}(x^k - \alpha_k \nabla f_{i(k)}(x^k))}$$

Choices of $i(k)$

- *Cyclic:* $i(k) = 1 + (k \mod m)$
- *Randomized:* Pick $i(k)$ uniformly from $\{1, \dots, m\}$

♣ Many other variations of incremental methods

♣ Read (omitting proofs) this nice survey by D. P. Bertsekas

Stochastic Optimization

Stochastic gradients

$$\boxed{\min f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x)}$$

Stochastic gradients

$$\boxed{\min f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x)}$$

Recall the incremental gradient method

- Let $x^0 \in \mathbb{R}^n$
- For $k \geq 0$

Stochastic gradients

$$\min f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x)$$

Recall the incremental gradient method

- Let $x^0 \in \mathbb{R}^n$
- For $k \geq 0$
  1. Pick $i(k) \in \{1, 2, \ldots, m\}$ uniformly at random
  2. $x^{k+1} = x^k - \alpha_k \nabla f_{i(k)}(x^k)$

Stochastic gradients

$$\min f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x)$$

Recall the incremental gradient method

- Let $x^0 \in \mathbb{R}^n$
- For $k \geq 0$
  1. Pick $i(k) \in \{1, 2, \ldots, m\}$ uniformly at random
  2. $x^{k+1} = x^k - \alpha_k \nabla f_{i(k)}(x^k)$

$g \equiv \nabla f_{i(k)}$ may be viewed as a **stochastic gradient**

Stochastic gradients

$$\boxed{\min f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x)}$$

Recall the incremental gradient method

- Let $x^0 \in \mathbb{R}^n$
- For $k \geq 0$
  1. Pick $i(k) \in \{1, 2, \ldots, m\}$ uniformly at random
  2. $x^{k+1} = x^k - \alpha_k \nabla f_{i(k)}(x^k)$

$\boxed{g \equiv \nabla f_{i(k)}}$ may be viewed as a **stochastic gradient**

$\boxed{g := g^{\text{true}} + \mathbf{e}}$, where e is mean-zero noise: $\mathbb{E}[e] = 0$

Stochastic gradients

- Index $i(k)$ chosen uniformly from $\{1, \ldots, m\}$
- Thus, **in expectation:**

$$\mathbb{E}[g] =$$

Stochastic gradients

- Index $i(k)$ chosen uniformly from $\{1, \ldots, m\}$
- Thus, **in expectation:**

$$\mathbb{E}[g] = \mathbb{E}_i[\nabla f_i(x)]$$

Stochastic gradients

- Index $i(k)$ chosen uniformly from $\{1, \ldots, m\}$
- Thus, **in expectation:**

$$\mathbb{E}[g] = \mathbb{E}_i[\nabla f_i(x)] = \sum\nolimits_i \frac{1}{m} \nabla f_i(x) =$$

Stochastic gradients

- Index $i(k)$ chosen uniformly from $\{1, \ldots, m\}$
- Thus, **in expectation:**

$$\mathbb{E}[g] = \mathbb{E}_i[\nabla f_i(x)] = \sum\nolimits_i \frac{1}{m} \nabla f_i(x) = \nabla \mathbf{f}(\mathbf{x})$$

Stochastic gradients

- Index $i(k)$ chosen uniformly from $\{1, \ldots, m\}$
- Thus, **in expectation:**

$$\mathbb{E}[g] = \mathbb{E}_i[\nabla f_i(x)] = \sum\nolimits_i \frac{1}{m} \nabla f_i(x) = \nabla \mathbf{f}(\mathbf{x})$$

- Alternatively, $\mathbb{E}[g - g^{\text{true}}] = \mathbb{E}[e] = 0$.

Stochastic gradients

- Index $i(k)$ chosen uniformly from $\{1, \ldots, m\}$
- Thus, **in expectation:**

$$\mathbb{E}[g] = \mathbb{E}_i[\nabla f_i(x)] = \sum_i \frac{1}{m} \nabla f_i(x) = \nabla \mathbf{f}(\mathbf{x})$$

- Alternatively, $\mathbb{E}[g - g^{\text{true}}] = \mathbb{E}[e] = 0$.
- We call g an **unbiased estimate** of the gradient

Stochastic gradients

- Index $i(k)$ chosen uniformly from $\{1, \ldots, m\}$
- Thus, **in expectation:**

$$\mathbb{E}[g] = \mathbb{E}_i[\nabla f_i(x)] = \sum_i \frac{1}{m} \nabla f_i(x) = \nabla \mathbf{f}(\mathbf{x})$$

- Alternatively, $\mathbb{E}[g - g^{\text{true}}] = \mathbb{E}[e] = 0$.
- We call g an **unbiased estimate** of the gradient
- Here, we **obtained** g in a two step process:
  - **Sample:** pick an index $i(k)$ unif. at random
  - **Oracle:** Compute a stochastic gradient based on $i(k)$

Stochastic programming

$$\boxed{\min f(x) := \mathbb{E}_\omega[F(x, \omega)]}$$

- ω follows some **known** distribution

Stochastic programming

$$\boxed{\min f(x) := \mathbb{E}_{\omega}[F(x, \omega)]}$$

- ω follows some **known** distribution
- Previous example, omega takes values in a **discrete set** of size m (might as well say $\omega \in \{1, \ldots, m\}$)

Stochastic programming

$$\boxed{\min f(x) := \mathbb{E}_\omega[F(x, \omega)]}$$

- ω follows some **known** distribution
- Previous example, omega takes values in a **discrete set** of size m (might as well say $\omega \in \{1, \ldots, m\}$)
- so that $F(x, \omega) = f_\omega(x)$; so assuming uniform distribution, we see that $f(x) = \mathbb{E}_\omega F(x, \omega) = \frac{1}{m}\sum_{i=1}^m f_i(x)$

Stochastic programming

$$\boxed{\min f(x) := \mathbb{E}_\omega[F(x,\omega)]}$$

- ω follows some **known** distribution
- Previous example, omega takes values in a **discrete set** of size m (might as well say $\omega \in \{1,\dots,m\}$)
- so that $F(x,\omega) = f_\omega(x)$; so assuming uniform distribution, we see that $f(x) = \mathbb{E}_\omega F(x,\omega) = \frac{1}{m}\sum_{i=1}^m f_i(x)$
- Usually ω will be **non-discrete**, and we won't be able to compute the expectation in closed form, since

  $$f(x) = \int F(x,\omega)dP(\omega),$$

  is going to be a difficult high-dimensional integral.

Stochastic programming – digression

Certainty-equivalent / mean approximation

- Say $F(x, \omega)$ is a linear function of x

Stochastic programming – digression

Certainty-equivalent / mean approximation

- Say $F(x, \omega)$ is a linear function of x
- Then, we may write $F(x, \omega) = \langle c(\omega),\, x \rangle$

Stochastic programming – digression

Certainty-equivalent / mean approximation

- Say $F(x, \omega)$ is a linear function of x
- Then, we may write $F(x, \omega) = \langle c(\omega), x \rangle$
- In this case, $f(x) = \int \langle c(\omega), x \rangle dP(\omega) = \langle \mathbb{E}[c(\omega)], x \rangle$

Stochastic programming – digression

Certainty-equivalent / mean approximation

- Say $F(x, \omega)$ is a linear function of x
- Then, we may write $F(x, \omega) = \langle c(\omega),\, x\rangle$
- In this case, $f(x) = \int \langle c(\omega),\, x\rangle dP(\omega) = \langle \mathbb{E}[c(\omega)],\, x\rangle$
- What if $F(x, \omega)$ is convex in x for every ω

Stochastic programming – digression

Certainty-equivalent / mean approximation

- Say $F(x, \omega)$ is a linear function of x
- Then, we may write $F(x, \omega) = \langle c(\omega), x \rangle$
- In this case, $f(x) = \int \langle c(\omega), x \rangle dP(\omega) = \langle \mathbb{E}[c(\omega)], x \rangle$
- What if $F(x, \omega)$ is convex in x for every ω
- Jensen's inequality gives us a trivial lower-bound

$$f(x) = \int F(x, \omega) dP(\omega) \geq F(x, \mathbb{E}[\omega])$$

Stochastic programming – digression

Certainty-equivalent / mean approximation

- Say $F(x, \omega)$ is a linear function of x
- Then, we may write $F(x, \omega) = \langle c(\omega),\, x\rangle$
- In this case, $f(x) = \int \langle c(\omega),\, x\rangle dP(\omega) = \langle \mathbb{E}[c(\omega)],\, x\rangle$
- What if $F(x, \omega)$ is convex in x for every ω
- Jensen's inequality gives us a trivial lower-bound

$$f(x) = \int F(x, \omega) dP(\omega) \geq F(x, \mathbb{E}[\omega])$$

- Bound may be too weak—even useless

Stochastic programming – digression

Certainty-equivalent / mean approximation

- Say $F(x,\omega)$ is a linear function of x
- Then, we may write $F(x,\omega) = \langle c(\omega),\, x\rangle$
- In this case, $f(x) = \int \langle c(\omega),\, x\rangle dP(\omega) = \langle \mathbb{E}[c(\omega)],\, x\rangle$
- What if $F(x,\omega)$ is convex in x for every ω
- Jensen's inequality gives us a trivial lower-bound

$$f(x) = \int F(x,\omega)dP(\omega) \geq F(x, \mathbb{E}[\omega])$$

- Bound may be too weak—even useless
- Thus, let us try to directly minimize $f(x)$

Stochastic programming – setup

$$\boxed{\min_{x \in \mathcal{X}} f(x) := \mathbb{E}_\omega[F(x,\omega)]}$$

Setup and Assumptions

1. $\mathcal{X} \subset \mathbb{R}^n$ nonempty, closed, bounded, convex

Stochastic programming – setup

$$\min_{x\in\mathcal{X}} f(x) := \mathbb{E}_\omega[F(x,\omega)]$$

Setup and Assumptions

1. $\mathcal{X} \subset \mathbb{R}^n$ nonempty, closed, bounded, convex
2. ω is a random vector whose probability distribution P is supported on $\Omega \subset \mathbb{R}^d$; so $F : \mathcal{X} \times \Omega \to \mathbb{R}$

Stochastic programming – setup

$$\boxed{\min_{x\in\mathcal{X}} f(x) := \mathbb{E}_\omega[F(x,\omega)]}$$

Setup and Assumptions

1. $\mathcal{X} \subset \mathbb{R}^n$ nonempty, closed, bounded, convex
2. ω is a random vector whose probability distribution P is supported on $\Omega \subset \mathbb{R}^d$; so $F : \mathcal{X} \times \Omega \to \mathbb{R}$
3. The expectation

$$\mathbb{E}[F(x,\omega)] = \int_\Omega F(x,\omega)dP(\omega)$$

is well-defined and finite valued for every $x \in \mathcal{X}$.

Stochastic programming – setup

$$\boxed{\min_{x\in\mathcal{X}} f(x) := \mathbb{E}_\omega[F(x,\omega)]}$$

Setup and Assumptions

1. $\mathcal{X} \subset \mathbb{R}^n$ nonempty, closed, bounded, convex

2. ω is a random vector whose probability distribution P is supported on $\Omega \subset \mathbb{R}^d$; so $F : \mathcal{X} \times \Omega \to \mathbb{R}$

3. The expectation

$$\mathbb{E}[F(x,\omega)] = \int_\Omega F(x,\omega)dP(\omega)$$

is well-defined and finite valued for every $x \in \mathcal{X}$.

4. For every $\omega \in \Omega$, $F(\cdot,\omega)$ is convex.

Convex stochastic optimization problem

Stochastic programming – setup

- Cannot compute expectation with high-accuracy in general

Stochastic programming – setup

- Cannot compute expectation with high-accuracy in general
- So, computational techniques based on Monte Carlo sampling

Stochastic programming – setup

- Cannot compute expectation with high-accuracy in general
- So, computational techniques based on Monte Carlo sampling

Assumption 1: Possible to generate independent identically distributed (iid) samples $\omega^1, \omega^2, \ldots$
Assumption 2: For a given input $(x, \omega) \in \mathcal{X} \times \Omega$, we can compute (oracle) a stochastic gradient $G(x, \omega)$

$$g(x) := \mathbb{E}[G(x, \omega)] \quad \text{s.t.} \quad g(x) \in \partial f(x).$$

Stochastic programming – setup

- Cannot compute expectation with high-accuracy in general
- So, computational techniques based on Monte Carlo sampling

Assumption 1: Possible to generate independent identically distributed (iid) samples $\omega^1, \omega^2, \dots$
Assumption 2: For a given input $(x, \omega) \in \mathcal{X} \times \Omega$, we can compute (oracle) a stochastic gradient $G(x, \omega)$

$$g(x) := \mathbb{E}[G(x, \omega)] \quad \text{s.t.} \quad g(x) \in \partial f(x).$$

- How to get these stochastic subgradients?

Stochastic programming – setup

- Cannot compute expectation with high-accuracy in general
- So, computational techniques based on Monte Carlo sampling

Assumption 1: Possible to generate independent identically distributed (iid) samples $\omega^1, \omega^2, \ldots$
Assumption 2: For a given input $(x, \omega) \in \mathcal{X} \times \Omega$, we can compute (oracle) a stochastic gradient $G(x, \omega)$

$$g(x) := \mathbb{E}[G(x, \omega)] \quad \text{s.t.} \quad g(x) \in \partial f(x).$$

- How to get these stochastic subgradients?

Theorem Let $\omega \in \Omega$; If $F(\cdot, \omega)$ is convex, and $f(\cdot)$ is finite valued in a neighborhood of a point x, then

$$\partial f(x) = \mathbb{E}[\partial_x F(x, \omega)].$$

Stochastic programming – setup

- Cannot compute expectation with high-accuracy in general
- So, computational techniques based on Monte Carlo sampling

Assumption 1: Possible to generate independent identically distributed (iid) samples $\omega^1, \omega^2, \ldots$
Assumption 2: For a given input $(x, \omega) \in \mathcal{X} \times \Omega$, we can compute (oracle) a stochastic gradient $G(x, \omega)$

$$g(x) := \mathbb{E}[G(x, \omega)] \quad \text{s.t.} \quad g(x) \in \partial f(x).$$

- How to get these stochastic subgradients?

Theorem Let $\omega \in \Omega$; If $F(\cdot, \omega)$ is convex, and $f(\cdot)$ is finite valued in a neighborhood of a point x, then

$$\partial f(x) = \mathbb{E}[\partial_x F(x, \omega)].$$

- So we may pick $G(x, \omega) \in \partial_x F(x, \omega)$ as stochastic subgradient.

Stochastic programming

- ♣ Stochastic Approximation (SA)
  - ▶ Sample ω^k iid

Stochastic programming

♣ Stochastic Approximation (SA)

- Sample ω^k iid
- Generate stochastic subgradient $G(x, \omega)$

Stochastic programming

- ♣ Stochastic Approximation (SA)
  - Sample ω^k iid
  - Generate stochastic subgradient $G(x, \omega)$
  - Use that in a subgradient method!

Stochastic programming

- Stochastic Approximation (SA)
  - Sample ω^k iid
  - Generate stochastic subgradient $G(x, \omega)$
  - Use that in a subgradient method!
- Sample average approximation (SAA)

Stochastic programming

- Stochastic Approximation (SA)
  - Sample ω^k iid
  - Generate stochastic subgradient $G(x, \omega)$
  - Use that in a subgradient method!
- Sample average approximation (SAA)
  - Generate N iid samples, $\omega^1, \ldots, \omega^N$

Stochastic programming

♣ Stochastic Approximation (SA)
- Sample ω^k iid
- Generate stochastic subgradient $G(x, \omega)$
- Use that in a subgradient method!

♣ Sample average approximation (SAA)
- Generate N iid samples, $\omega^1, \ldots, \omega^N$
- Consider **empirical objective** $\hat{f}_N := N^{-1} \sum_i F(x, \omega^i)$

Stochastic programming

- ♣ Stochastic Approximation (SA)
  - Sample ω^k iid
  - Generate stochastic subgradient $G(x, \omega)$
  - Use that in a subgradient method!
- ♣ Sample average approximation (SAA)
  - Generate N iid samples, $\omega^1, \ldots, \omega^N$
  - Consider **empirical objective** $\hat{f}_N := N^{-1} \sum_i F(x, \omega^i)$
  - SAA refers to creation of this **sample average problem**
  - Minimizing $\hat{f}_N$ still needs to be done!

Stochastic approximation – SA

SA or stochastic (sub)-gradient

- Let $x^0 \in \mathcal{X}$
- For $k \geq 0$
  - Sample ω^k iid; generate $G(x^k, \omega^k)$
  - Update $x^{k+1} = P_{\mathcal{X}}(x^k - \alpha_k G(x^k, \omega^k))$, where $\alpha_k > 0$

Stochastic approximation – SA

SA or stochastic (sub)-gradient

- Let $x^0 \in \mathcal{X}$
- For $k \geq 0$
  - Sample ω^k iid; generate $G(x^k, \omega^k)$
  - Update $x^{k+1} = P_{\mathcal{X}}(x^k - \alpha_k G(x^k, \omega^k))$, where $\alpha_k > 0$

Henceforth, we'll simply write:

$$\boxed{x^{k+1} = P_{\mathcal{X}}\left(x^k - \alpha_k G^k\right)}$$

Stochastic approximation – SA

SA or stochastic (sub)-gradient

- Let $x^0 \in \mathcal{X}$
- For $k \geq 0$
  - Sample ω^k iid; generate $G(x^k, \omega^k)$
  - Update $x^{k+1} = P_{\mathcal{X}}(x^k - \alpha_k G(x^k, \omega^k))$, where $\alpha_k > 0$

Henceforth, we'll simply write:

$$\boxed{x^{k+1} = P_{\mathcal{X}}\left(x^k - \alpha_k G^k\right)}$$

Does this work?

Stochastic approximation – analysis

Setup

- x^k depends on rvs $\omega^1, \ldots, \omega^{k-1}$, so itself random

Stochastic approximation – analysis

Setup

- x^k depends on rvs $\omega^1, \ldots, \omega^{k-1}$, so itself random
- Of course, x^k does not depend on ω^k

Stochastic approximation – analysis

Setup

- x^k depends on rvs $\omega^1, \ldots, \omega^{k-1}$, so itself random
- Of course, x^k does not depend on ω^k
- Subgradient method analysis hinged upon: $\|x^k - x^*\|_2^2$

Stochastic approximation – analysis

Setup

- x^k depends on rvs $\omega^1, \ldots, \omega^{k-1}$, so itself random
- Of course, x^k does not depend on ω^k
- Subgradient method analysis hinged upon: $\|x^k - x^*\|_2^2$
- Stochastic subgradient hinges upon: $\mathbb{E}[\|x^k - x^*\|_2^2]$

Stochastic approximation – analysis

Setup

- x^k depends on rvs $\omega^1, \ldots, \omega^{k-1}$, so itself random
- Of course, x^k does not depend on ω^k
- Subgradient method analysis hinged upon: $\|x^k - x^*\|_2^2$
- Stochastic subgradient hinges upon: $\mathbb{E}[\|x^k - x^*\|_2^2]$

Denote: $R_k := \|x^k - x^*\|_2^2$ and $r_k := \mathbb{E}[R_k] = \mathbb{E}[\|x^k - x^*\|_2^2]$

Stochastic approximation – analysis

Setup

- x^k depends on rvs $\omega^1, \ldots, \omega^{k-1}$, so itself random
- Of course, x^k does not depend on ω^k
- Subgradient method analysis hinged upon: $\|x^k - x^*\|_2^2$
- Stochastic subgradient hinges upon: $\mathbb{E}[\|x^k - x^*\|_2^2]$

Denote: $R_k := \|x^k - x^*\|_2^2$ and $r_k := \mathbb{E}[R_k] = \mathbb{E}[\|x^k - x^*\|_2^2]$

Bounding R_{k+1}

$$\begin{aligned}
R_{k+1} &= \|x^{k+1} - x^*\|_2^2 = \|P_{\mathcal{X}}(x^k - \alpha_k G^k) - P_{\mathcal{X}} x^*\|_2^2 \\
&\leq \|x^k - x^* - \alpha_k G^k\|_2^2 \\
&= R_k + \alpha_k^2 \|G^k\|_2^2 - 2\alpha_k \langle G^k,\, x^k - x^* \rangle.
\end{aligned}$$

Stochastic approximation – analysis

$$\boxed{R_{k+1} \leq R_k + \alpha_k^2 \|G^k\|_2^2 - 2\alpha_k \langle G^k,\, x^k - x^* \rangle}$$

Stochastic approximation – analysis

$$\boxed{R_{k+1} \leq R_k + \alpha_k^2 \|G^k\|_2^2 - 2\alpha_k \langle G^k,\, x^k - x^* \rangle}$$

- **Assume:** $\|G^k\|_2 \leq M$ on $\mathcal{X}$
- Taking expectation:

$$r_{k+1} \leq r_k + \alpha_k^2 M^2 - 2\alpha_k \mathbb{E}[\langle G^k,\, x^k - x^* \rangle].$$

Stochastic approximation – analysis

$$\boxed{R_{k+1} \le R_k + \alpha_k^2 \|G^k\|_2^2 - 2\alpha_k \langle G^k, x^k - x^* \rangle}$$

- **Assume:** $\|G^k\|_2 \le M$ on $\mathcal{X}$
- Taking expectation:
$$r_{k+1} \le r_k + \alpha_k^2 M^2 - 2\alpha_k \mathbb{E}[\langle G^k, x^k - x^* \rangle].$$
- We need to now get a handle on the last term

Stochastic approximation – analysis

$$\boxed{R_{k+1} \leq R_k + \alpha_k^2 \|G^k\|_2^2 - 2\alpha_k \langle G^k, x^k - x^* \rangle}$$

- **Assume:** $\|G^k\|_2 \leq M$ on $\mathcal{X}$
- Taking expectation:
$$r_{k+1} \leq r_k + \alpha_k^2 M^2 - 2\alpha_k \mathbb{E}[\langle G^k, x^k - x^* \rangle].$$
- We need to now get a handle on the last term
- Since x^k is independent of ω^k, we have

$$\mathbb{E}[\langle x^k - x^*, G(x^k, \omega^k) \rangle] \quad =$$

Stochastic approximation – analysis

$$\boxed{R_{k+1} \leq R_k + \alpha_k^2 \|G^k\|_2^2 - 2\alpha_k \langle G^k,\, x^k - x^* \rangle}$$

- **Assume:** $\|G^k\|_2 \leq M$ on $\mathcal{X}$
- Taking expectation:

$$r_{k+1} \leq r_k + \alpha_k^2 M^2 - 2\alpha_k \mathbb{E}[\langle G^k,\, x^k - x^* \rangle].$$

- We need to now get a handle on the last term
- Since x^k is independent of ω^k, we have

$$\begin{aligned} \mathbb{E}[\langle x^k - x^*,\, G(x^k, \omega^k) \rangle] &= \mathbb{E}\left\{ \mathbb{E}[\langle x^k - x^*,\, G(x^k, \omega^k) \rangle \mid \omega^{1..(k-1)}] \right\} \\ &= \end{aligned}$$

Stochastic approximation – analysis

$$\boxed{R_{k+1} \le R_k + \alpha_k^2 \|G^k\|_2^2 - 2\alpha_k \langle G^k,\, x^k - x^* \rangle}$$

- **Assume:** $\|G^k\|_2 \le M$ on $\mathcal{X}$
- Taking expectation:

$$r_{k+1} \le r_k + \alpha_k^2 M^2 - 2\alpha_k \mathbb{E}[\langle G^k,\, x^k - x^* \rangle].$$

- We need to now get a handle on the last term
- Since x^k is independent of ω^k, we have

$$\begin{aligned}
\mathbb{E}[\langle x^k - x^*,\, G(x^k, \omega^k)\rangle] &= \mathbb{E}\left\{\mathbb{E}[\langle x^k - x^*,\, G(x^k, \omega^k)\rangle \mid \omega^{1..(k-1)}]\right\} \\
&= \mathbb{E}\left\{\langle x^k - x^*,\, \mathbb{E}[G(x^k, \omega^k) \mid \omega^{1..k-1}]\rangle\right\} \\
&=
\end{aligned}$$

Stochastic approximation – analysis

$$\boxed{R_{k+1} \leq R_k + \alpha_k^2 \|G^k\|_2^2 - 2\alpha_k \langle G^k,\, x^k - x^* \rangle}$$

- **Assume:** $\|G^k\|_2 \leq M$ on $\mathcal{X}$
- Taking expectation:

$$r_{k+1} \leq r_k + \alpha_k^2 M^2 - 2\alpha_k \mathbb{E}[\langle G^k,\, x^k - x^* \rangle].$$

- We need to now get a handle on the last term
- Since x^k is independent of ω^k, we have

$$\begin{aligned}
\mathbb{E}[\langle x^k - x^*,\, G(x^k, \omega^k) \rangle] &= \mathbb{E}\left\{ \mathbb{E}[\langle x^k - x^*,\, G(x^k, \omega^k) \rangle \mid \omega^{1..(k-1)}] \right\} \\
&= \mathbb{E}\left\{ \langle x^k - x^*,\, \mathbb{E}[G(x^k, \omega^k) \mid \omega^{1..k-1}] \rangle \right\} \\
&= \mathbb{E}[\langle x^k - x^*,\, g^k \rangle], \quad g^k \in \partial f(x^k).
\end{aligned}$$

Stochastic approximation – analysis

$$\boxed{\text{Thus, we need to bound: } \mathbb{E}[\langle x^k - x^*, g^k \rangle]}$$

Stochastic approximation – analysis

$$\boxed{\text{Thus, we need to bound: } \mathbb{E}[\langle x^k - x^*,\, g^k \rangle]}$$

- Since f is cvx, $f(x) \geq f(x^k) + \langle g^k,\, x - x^k \rangle$ for any $x \in \mathcal{X}$.

Stochastic approximation – analysis

$$\boxed{\text{Thus, we need to bound: } \mathbb{E}[\langle x^k - x^*,\, g^k \rangle]}$$

- Since f is cvx, $f(x) \geq f(x^k) + \langle g^k,\, x - x^k \rangle$ for any $x \in \mathcal{X}$.
- Thus, in particular we have

$$2\alpha_k \mathbb{E}[f(x^*) - f(x^k)] \geq 2\alpha_k \mathbb{E}[\langle g^k,\, x^* - x^k \rangle]$$

Stochastic approximation – analysis

$$\boxed{\text{Thus, we need to bound: } \mathbb{E}[\langle x^k - x^*, g^k \rangle]}$$

- Since f is cvx, $f(x) \geq f(x^k) + \langle g^k, x - x^k \rangle$ for any $x \in \mathcal{X}$.
- Thus, in particular we have

$$2\alpha_k \mathbb{E}[f(x^*) - f(x^k)] \geq 2\alpha_k \mathbb{E}[\langle g^k, x^* - x^k \rangle]$$

Now plug this bound back into the r_{k+1} inequality

$$r_{k+1} \leq r_k + \alpha_k^2 M^2 - 2\alpha_k \mathbb{E}[\langle g^k, x^k - x^* \rangle]$$

Stochastic approximation – analysis

$$\boxed{\text{Thus, we need to bound: } \mathbb{E}[\langle x^k - x^*,\, g^k\rangle]}$$

- Since f is cvx, $f(x) \geq f(x^k) + \langle g^k,\, x - x^k\rangle$ for any $x \in \mathcal{X}$.
- Thus, in particular we have

$$2\alpha_k \mathbb{E}[f(x^*) - f(x^k)] \geq 2\alpha_k \mathbb{E}[\langle g^k,\, x^* - x^k\rangle]$$

Now plug this bound back into the r_{k+1} inequality

$$\begin{aligned} r_{k+1} &\leq r_k + \alpha_k^2 M^2 - 2\alpha_k \mathbb{E}[\langle g^k,\, x^k - x^*\rangle] \\ 2\alpha_k \mathbb{E}[\langle g^k,\, x^k - x^*\rangle] &\leq r_k - r_{k+1} + \alpha_k M^2 \end{aligned}$$

Stochastic approximation – analysis

Thus, we need to bound: $\mathbb{E}[\langle x^k - x^*, g^k\rangle]$

- Since f is cvx, $f(x) \geq f(x^k) + \langle g^k, x - x^k\rangle$ for any $x \in \mathcal{X}$.
- Thus, in particular we have

$$2\alpha_k \mathbb{E}[f(x^*) - f(x^k)] \geq 2\alpha_k \mathbb{E}[\langle g^k, x^* - x^k\rangle]$$

Now plug this bound back into the r_{k+1} inequality

$$\begin{aligned} r_{k+1} &\leq r_k + \alpha_k^2 M^2 - 2\alpha_k \mathbb{E}[\langle g^k, x^k - x^*\rangle] \\ 2\alpha_k \mathbb{E}[\langle g^k, x^k - x^*\rangle] &\leq r_k - r_{k+1} + \alpha_k M^2 \\ 2\alpha_k \mathbb{E}[f(x^k) - f(x^*)] &\leq r_k - r_{k+1} + \alpha_k M^2. \end{aligned}$$

Stochastic approximation – analysis

Thus, we need to bound: $\mathbb{E}[\langle x^k - x^*, g^k \rangle]$

- Since f is cvx, $f(x) \geq f(x^k) + \langle g^k, x - x^k \rangle$ for any $x \in \mathcal{X}$.
- Thus, in particular we have

$$2\alpha_k \mathbb{E}[f(x^*) - f(x^k)] \geq 2\alpha_k \mathbb{E}[\langle g^k, x^* - x^k \rangle]$$

Now plug this bound back into the r_{k+1} inequality

$$\begin{aligned} r_{k+1} &\leq r_k + \alpha_k^2 M^2 - 2\alpha_k \mathbb{E}[\langle g^k, x^k - x^* \rangle] \\ 2\alpha_k \mathbb{E}[\langle g^k, x^k - x^* \rangle] &\leq r_k - r_{k+1} + \alpha_k M^2 \\ 2\alpha_k \mathbb{E}[f(x^k) - f(x^*)] &\leq r_k - r_{k+1} + \alpha_k M^2. \end{aligned}$$

What now?

Stochastic approximation – analysis

$$2\alpha_k \mathbb{E}[f(x^k) - f(x^*)] \leq r_k - r_{k+1} + \alpha_k M^2.$$

Stochastic approximation – analysis

$$2\alpha_k \mathbb{E}[f(x^k) - f(x^*)] \leq r_k - r_{k+1} + \alpha_k M^2.$$

Sum up over $k = 1, \ldots, T$, to obtain

$$\sum\nolimits_{k=1}^{T} (2\alpha_k \mathbb{E}[f(x^k) - f(x^*)]) \quad \leq \quad r_1 - r_{T+1} + M^2 \sum\nolimits_k \alpha_k^2$$

Stochastic approximation – analysis

$$2\alpha_k \mathbb{E}[f(x^k) - f(x^*)] \le r_k - r_{k+1} + \alpha_k M^2.$$

Sum up over $k = 1, \ldots, T$, to obtain

$$\begin{aligned}
\sum\nolimits_{k=1}^{T} (2\alpha_k \mathbb{E}[f(x^k) - f(x^*)]) &\le r_1 - r_{T+1} + M^2 \sum\nolimits_k \alpha_k^2 \\
&\le r_1 + M^2 \sum\nolimits_k \alpha_k^2.
\end{aligned}$$

Stochastic approximation – analysis

$$2\alpha_k \mathbb{E}[f(x^k) - f(x^*)] \le r_k - r_{k+1} + \alpha_k M^2.$$

Sum up over $k = 1, \ldots, T$, to obtain

$$\begin{aligned} \sum\nolimits_{k=1}^{T} (2\alpha_k \mathbb{E}[f(x^k) - f(x^*)]) &\le r_1 - r_{T+1} + M^2 \sum\nolimits_k \alpha_k^2 \\ &\le r_1 + M^2 \sum\nolimits_k \alpha_k^2. \end{aligned}$$

To further analyze this sum, divide both sides by $\sum_k \alpha_k$, so

Stochastic approximation – analysis

$$2\alpha_k \mathbb{E}[f(x^k) - f(x^*)] \leq r_k - r_{k+1} + \alpha_k M^2.$$

Sum up over $k = 1, \ldots, T$, to obtain

$$\begin{aligned} \sum\nolimits_{k=1}^{T} (2\alpha_k \mathbb{E}[f(x^k) - f(x^*)]) &\leq r_1 - r_{T+1} + M^2 \sum\nolimits_k \alpha_k^2 \\ &\leq r_1 + M^2 \sum\nolimits_k \alpha_k^2. \end{aligned}$$

To further analyze this sum, divide both sides by $\sum_k \alpha_k$, so

- Set $\gamma_k = \frac{\alpha_k}{\sum_k^T \alpha_k}$.
- Thus, $\gamma_k \geq 0$ and $\sum_k \gamma_k = 1$; this allows us to write

Stochastic approximation – analysis

$$2\alpha_k \mathbb{E}[f(x^k) - f(x^*)] \leq r_k - r_{k+1} + \alpha_k M^2.$$

Sum up over $k = 1, \ldots, T$, to obtain

$$\begin{aligned} \sum\nolimits_{k=1}^{T} (2\alpha_k \mathbb{E}[f(x^k) - f(x^*)]) &\leq r_1 - r_{T+1} + M^2 \sum\nolimits_k \alpha_k^2 \\ &\leq r_1 + M^2 \sum\nolimits_k \alpha_k^2. \end{aligned}$$

To further analyze this sum, divide both sides by $\sum_k \alpha_k$, so

- Set $\gamma_k = \frac{\alpha_k}{\sum_k^T \alpha_k}$.
- Thus, $\gamma_k \geq 0$ and $\sum_k \gamma_k = 1$; this allows us to write

$$\mathbb{E}\left[\sum\nolimits_k \gamma_k f(x^k) - f(x^*)\right] \leq \frac{r_1 + M^2 \sum_k \alpha_k^2}{2 \sum_k \alpha_k}$$

Stochastic approximation – analysis

- Bound looks similar to bound in subgradient method!

Stochastic approximation – analysis

- Bound looks similar to bound in subgradient method!
- But we wish to say something about x^T

Stochastic approximation – analysis

- Bound looks similar to bound in subgradient method!
- But we wish to say something about x^T
- Since $\gamma_k \geq 0$ and $\sum_k \gamma_k = 1$, and we have $\gamma_k f(x^k)$

Stochastic approximation – analysis

- Bound looks similar to bound in subgradient method!
- But we wish to say something about x^T
- Since $\gamma_k \geq 0$ and $\sum_k \gamma_k = 1$, and we have $\gamma_k f(x^k)$
- Easier to talk about **average iterate**

$$x_{av}^T := \sum_k^T \gamma_k x^k.$$

Stochastic approximation – analysis

- Bound looks similar to bound in subgradient method!
- But we wish to say something about x^T
- Since $\gamma_k \geq 0$ and $\sum_k \gamma_k = 1$, and we have $\gamma_k f(x^k)$
- Easier to talk about **average iterate**

$$x_{av}^T := \sum_k^T \gamma_k x^k.$$

- $f(x_{av}^T) \leq \sum_m \gamma_k f(x^k)$ due to convexity

Stochastic approximation – analysis

- Bound looks similar to bound in subgradient method!
- But we wish to say something about x^T
- Since $\gamma_k \geq 0$ and $\sum_k \gamma_k = 1$, and we have $\gamma_k f(x^k)$
- Easier to talk about **average iterate**

$$x_{av}^T := \sum_k^T \gamma_k x^k.$$

- $f(x_{av}^T) \leq \sum_m \gamma_k f(x^k)$ due to convexity
- So we finally obtain the inequality

$$\mathbb{E}\big[f(x_{av}) - f(x^*)\big] \leq \frac{r_1 + M^2 \sum_k \alpha_k^2}{2 \sum_k \alpha_k}.$$

Stochastic approximation – analysis

Exercise

- ♠ Let $D_{\mathcal{X}} := \max_{x \in \mathcal{X}} \|x - x^*\|_2$
- ♠ Assume $\alpha_k = \alpha$ is a constant. Then, observe that

$$\mathbb{E}[f(x_{av}^T) - f(x^*)] \leq \frac{D_{\mathcal{X}}^2 + M^2 T \alpha^2}{2T\alpha}$$

- ♠ Minimize the rhs over $\alpha > 0$ to obtain the best stepsize
- ♠ Show that this choice then yields: $\mathbb{E}[f(x_{av}^T) - f(x^*)] \leq \frac{D_{\mathcal{X}} M}{\sqrt{T}}$
- ♠ If T is not fixed in advance, then choose

$$\alpha_k = \frac{\theta D_{\mathcal{X}}}{M\sqrt{k}}, \quad k = 1, 2, \ldots$$

- ♠ Analyze $\mathbb{E}[f(x_{av}^T) - f(x^*)]$ with this choice of stepsize

Sample average approximation

Assumption: regularization $\|x\|_2 \leq B$; $\omega \in \Omega$ closed, bounded.

Function estimate: $f(x) = \mathbb{E}[F(x, \omega)]$
Subgradient in $\partial f(x) = \mathbb{E}[G(x, \omega)]$

Sample Average Approximation (SAA):

- Collect samples $\omega^1, \ldots, \omega^N$
- Empirical objective: $\hat{f}_N(x) := \frac{1}{N} \sum_{i=1}^N F(x, \omega^i)$

Sample average approximation

Assumption: regularization $\|x\|_2 \leq B$; $\omega \in \Omega$ closed, bounded.

$$\text{Function estimate: } f(x) = \mathbb{E}[F(x,\omega)]$$
$$\text{Subgradient in } \partial f(x) = \mathbb{E}[G(x,\omega)]$$

Sample Average Approximation (SAA):

- Collect samples $\omega^1, \ldots, \omega^N$
- Empirical objective: $\hat{f}_N(x) := \frac{1}{N}\sum_{i=1}^N F(x,\omega^i)$
- aka *Empirical Risk Minimization*

Sample average approximation

Assumption: regularization $\|x\|_2 \leq B$; $\omega \in \Omega$ closed, bounded.

Function estimate: $f(x) = \mathbb{E}[F(x,\omega)]$
Subgradient in $\partial f(x) = \mathbb{E}[G(x,\omega)]$

Sample Average Approximation (SAA):

- Collect samples $\omega^1, \ldots, \omega^N$
- Empirical objective: $\hat{f}_N(x) := \frac{1}{N}\sum_{i=1}^N F(x,\omega^i)$
- aka *Empirical Risk Minimization*
- Confusing: Machine learners often optimize $\hat{f}_N$ using stochastic subgradient; but theoretical guarantees are then only on the *empirical* suboptimality $E[\hat{f}_N(\bar{x}^k)] \leq \ldots$

Sample average approximation

Assumption: regularization $\|x\|_2 \leq B$; $\omega \in \Omega$ closed, bounded.

Function estimate: $f(x) = \mathbb{E}[F(x,\omega)]$
Subgradient in $\partial f(x) = \mathbb{E}[G(x,\omega)]$

Sample Average Approximation (SAA):

- Collect samples $\omega^1, \ldots, \omega^N$
- Empirical objective: $\hat{f}_N(x) := \frac{1}{N}\sum_{i=1}^N F(x,\omega^i)$
- aka *Empirical Risk Minimization*
- Confusing: Machine learners often optimize $\hat{f}_N$ using stochastic subgradient; but theoretical guarantees are then only on the *empirical* suboptimality $E[\hat{f}_N(\bar{x}^k)] \leq \ldots$
- For guarantees on $f(\bar{x}^k)$, extra work is needed *regularization* + unif. concentration used
$f(\bar{x}^k) - f(x^*) \leq O(1/\sqrt{k}) + O(1/\sqrt{N})$

Stochastic Programming – modeling

Stochastic LP

$$
\begin{aligned}
\min \quad & x_1 + x_2 \\
\omega_1 x_1 + x_2 &\geq 10 \\
\omega_2 x_1 + x_2 &\geq 5 \\
x_1, x_2 &\geq 0,
\end{aligned}
$$

where $\omega_1 \sim \mathcal{U}[1,5]$ and $\omega_2 \sim \mathcal{U}[1/3,1]$

Stochastic Programming – modeling

Stochastic LP

$$
\begin{array}{rcl}
\min & & x_1 + x_2 \\
\omega_1 x_1 + x_2 & \geq & 10 \\
\omega_2 x_1 + x_2 & \geq & 5 \\
x_1, x_2 & \geq & 0,
\end{array}
$$

where $\omega_1 \sim \mathcal{U}[1,5]$ and $\omega_2 \sim \mathcal{U}[1/3,1]$

- The constraints are not deterministic!
- But we have an idea about what randomness is there

Stochastic Programming – modeling

Stochastic LP

$$\begin{array}{rcl} \min & & x_1 + x_2 \\ \omega_1 x_1 + x_2 & \geq & 10 \\ \omega_2 x_1 + x_2 & \geq & 5 \\ x_1, x_2 & \geq & 0, \end{array}$$

where $\omega_1 \sim \mathcal{U}[1,5]$ and $\omega_2 \sim \mathcal{U}[1/3,1]$

- The constraints are not deterministic!
- But we have an idea about what randomness is there
- How do we *solve* this LP?

Stochastic Programming – modeling

Stochastic LP

$$
\begin{aligned}
\min \quad & x_1 + x_2 \\
\omega_1 x_1 + x_2 &\geq 10 \\
\omega_2 x_1 + x_2 &\geq 5 \\
x_1, x_2 &\geq 0,
\end{aligned}
$$

where $\omega_1 \sim \mathcal{U}[1,5]$ and $\omega_2 \sim \mathcal{U}[1/3,1]$

- The constraints are not deterministic!
- But we have an idea about what randomness is there
- How do we *solve* this LP?
- What does it even mean to solve it?

Stochastic Programming – modeling

Stochastic LP

$$\begin{aligned} \min \quad & x_1 + x_2 \\ \omega_1 x_1 + x_2 &\geq 10 \\ \omega_2 x_1 + x_2 &\geq 5 \\ x_1, x_2 &\geq 0, \end{aligned}$$

where $\omega_1 \sim \mathcal{U}[1,5]$ and $\omega_2 \sim \mathcal{U}[1/3,1]$

- The constraints are not deterministic!
- But we have an idea about what randomness is there
- How do we *solve* this LP?
- What does it even mean to solve it?
- If ω **has been observed**, problem becomes deterministic, and can be solved as a usual LP (aka **wait-and-watch**)

Stochastic Programming – modeling

- But we cannot "wait-and-watch" —

Stochastic Programming – modeling

- But we cannot "wait-and-watch" — we need to decide on x *before knowing* the value of ω

Stochastic Programming – modeling

- But we cannot "wait-and-watch" — we need to decide on x *before knowing* the value of ω
- What to do without knowing exact values for ω_1, ω_2?

Stochastic Programming – modeling

- But we cannot "wait-and-watch" — we need to decide on x *before knowing* the value of ω
- What to do without knowing exact values for ω_1, ω_2?
- Some ideas
  - Guess the uncertainty
  - Probabilistic / Chance constraints
  - ...

Stochastic Programming – modeling

Some guesses

- ♠ *Unbiased / Average case:* Choose **mean values** for each r.v.
- ♠ *Robust / Worst case:* Choose **worst case** values
- ♠ *Explorative / Best case:* Choose **best case** values

Stochastic Programming – Example

$$\begin{aligned} \min \quad & x_1 + x_2 \\ \omega_1 x_1 + x_2 \quad & \geq \quad 10 \\ \omega_2 x_1 + x_2 \quad & \geq \quad 5 \\ x_1, x_2 \quad & \geq \quad 0, \end{aligned}$$

where $\omega_1 \sim \mathcal{U}[1,5]$ and $\omega_2 \sim \mathcal{U}[1/3,1]$

Unbiased / Average case:

$$\mathbb{E}[\omega_1] = 3, \quad \mathbb{E}[\omega_2] = 2/3$$

$$\begin{aligned} \min \quad & x_1 + x_2 \\ 3x_1 + x_2 \quad & \geq \quad 10 \\ (2/3)x_1 + x_2 \quad & \geq \quad 5 \\ x_1, x_2 \quad & \geq \quad 0, \end{aligned}$$

$x_1^* + x_2^* = \mathbf{5.7143...}$

$(x_1^*, x_2^*) \approx (15/7, 25/7).$

Stochastic Programming – Example

$$\begin{aligned} \min \quad & x_1 + x_2 \\ \omega_1 x_1 + x_2 & \geq 10 \\ \omega_2 x_1 + x_2 & \geq 5 \\ x_1, x_2 & \geq 0, \end{aligned}$$

where $\omega_1 \sim \mathcal{U}[1,5]$ and $\omega_2 \sim \mathcal{U}[1/3,1]$

Worst case:

$$\mathbb{E}[\omega_1] = 3, \quad \mathbb{E}[\omega_2] = 2/3$$

$$\begin{aligned} \min \quad & x_1 + x_2 \\ 1 x_1 + x_2 & \geq 10 \\ (1/3) x_1 + x_2 & \geq 5 \\ x_1, x_2 & \geq 0, \end{aligned}$$

$$x_1^* + x_2^* = \mathbf{10}$$

$$(x_1^*, x_2^*) \approx (41/12, 79/12).$$

Stochastic Programming – Example

$$\begin{aligned} \min \quad & x_1 + x_2 \\ \omega_1 x_1 + x_2 &\geq 10 \\ \omega_2 x_1 + x_2 &\geq 5 \\ x_1, x_2 &\geq 0, \end{aligned}$$

where $\omega_1 \sim \mathcal{U}[1,5]$ and $\omega_2 \sim \mathcal{U}[1/3,1]$

Best case:

$$\mathbb{E}[\omega_1] = 3, \quad \mathbb{E}[\omega_2] = 2/3$$

$$\begin{aligned} \min \quad & x_1 + x_2 \\ 5x_1 + x_2 &\geq 10 \\ 1x_1 + x_2 &\geq 5 \\ x_1, x_2 &\geq 0, \end{aligned}$$

$x_1^* + x_2^* = \mathbf{5}$

$(x_1^*, x_2^*) \approx (17/8, 23/8).$

Online optimization

Online optimization

- We have *fixed* and *known* $F(x, \omega)$

Online optimization

- We have *fixed* and *known* $F(x, \omega)$
- $\omega^1, \omega^2, \ldots$ presented to us sequentially

Can be chosen adversarially!

Online optimization

- We have *fixed* and *known* $F(x, \omega)$
- $\omega^1, \omega^2, \ldots$ presented to us sequentially

Can be chosen adversarially!

- **Guess** x^k;

Online optimization

- We have *fixed* and *known* $F(x, \omega)$
- $\omega^1, \omega^2, \ldots$ presented to us sequentially

Can be chosen adversarially!

- **Guess** x^k; **Observe** ω^k;

Online optimization

- We have *fixed* and *known* $F(x, \omega)$
- $\omega^1, \omega^2, \ldots$ presented to us sequentially

Can be chosen adversarially!

- **Guess** x^k; **Observe** ω^k; **incur cost** $F(x^k, \omega^k)$;

Online optimization

- We have *fixed* and *known* $F(x, \omega)$
- $\omega^1, \omega^2, \ldots$ presented to us sequentially

Can be chosen adversarially!

- **Guess** x^k; **Observe** ω^k; **incur cost** $F(x^k, \omega^k)$; **Update** to x^{k+1}

Online optimization

- We have *fixed* and *known* $F(x, \omega)$
- $\omega^1, \omega^2, \ldots$ presented to us sequentially

Can be chosen adversarially!

- **Guess** x^k; **Observe** ω^k; **incur cost** $F(x^k, \omega^k)$; **Update** to x^{k+1}
- We get to see things only sequentially, and the sequence of samples shown to us by nature may depend on our guesses

Online optimization

- We have *fixed* and *known* $F(x, \omega)$
- $\omega^1, \omega^2, \ldots$ presented to us sequentially

$$\boxed{\textbf{Can be chosen adversarially!}}$$

- **Guess** x^k; **Observe** ω^k; **incur cost** $F(x^k, \omega^k)$; **Update** to x^{k+1}
- We get to see things only sequentially, and the sequence of samples shown to us by nature may depend on our guesses
- So a typical goal is to minimize **Regret**

Online optimization

- We have *fixed* and *known* $F(x, \omega)$
- $\omega^1, \omega^2, \ldots$ presented to us sequentially

Can be chosen adversarially!

- **Guess** x^k; **Observe** ω^k; **incur cost** $F(x^k, \omega^k)$; **Update** to x^{k+1}
- We get to see things only sequentially, and the sequence of samples shown to us by nature may depend on our guesses
- So a typical goal is to minimize **Regret**

$$\frac{1}{T}\sum_{k=1}^{T} F(x_k, z_k) - \min_{x\in\mathcal{X}} \frac{1}{T}\sum_{k=1}^{T} F(x, z_k)$$

Online optimization

- We have *fixed* and *known* $F(x, \omega)$
- $\omega^1, \omega^2, \ldots$ presented to us sequentially

Can be chosen adversarially!

- **Guess** x^k; **Observe** ω^k; **incur cost** $F(x^k, \omega^k)$; **Update** to x^{k+1}
- We get to see things only sequentially, and the sequence of samples shown to us by nature may depend on our guesses
- So a typical goal is to minimize **Regret**

$$\frac{1}{T}\sum_{k=1}^{T} F(x_k, z_k) - \min_{x \in \mathcal{X}} \frac{1}{T}\sum_{k=1}^{T} F(x, z_k)$$

- That is, difference from the best possible solution we could have attained, had we been shown all the examples (z_k).

Online optimization

- We have *fixed* and *known* $F(x, \omega)$
- $\omega^1, \omega^2, \ldots$ presented to us sequentially

$$\boxed{\textbf{Can be chosen adversarially!}}$$

- **Guess** x^k; **Observe** ω^k; **incur cost** $F(x^k, \omega^k)$; **Update** to x^{k+1}
- We get to see things only sequentially, and the sequence of samples shown to us by nature may depend on our guesses
- So a typical goal is to minimize **Regret**

$$\boxed{\frac{1}{T}\sum_{k=1}^{T} F(x_k, z_k) - \min_{x\in\mathcal{X}} \frac{1}{T}\sum_{k=1}^{T} F(x, z_k)}$$

- That is, difference from the best possible solution we could have attained, had we been shown all the examples (z_k).
- Online optimization is an important idea in machine learning, game theory, decision making, etc.

Online gradient descent

Based on Zinkevich (2003)

Slight generalization:
$F(x, \omega)$ convex (in x); possibly nonsmooth
$x \in \mathcal{X}$, a closed, bounded set

Online gradient descent

Based on Zinkevich (2003)

Slight generalization:
$F(x, \omega)$ convex (in x); possibly nonsmooth
$x \in \mathcal{X}$, a closed, bounded set

Simplify notation: $f_k(x) \equiv F(x, \omega^k)$

Regret $R_T := \sum_{k=1}^{T} f_k(x^k) - \min_{x \in \mathcal{X}} \sum_{k=1}^{T} f_k(x)$

Online gradient descent

Algorithm:

1. Select some $x^0 \in \mathcal{X}$, and $\alpha_0 > 0$
2. Round k of algo ($k \geq 0$):

Online gradient descent

Algorithm:

1. Select some $x^0 \in \mathcal{X}$, and $\alpha_0 > 0$
2. Round k of algo ($k \geq 0$):
   - Output x^k

Online gradient descent

Algorithm:

1. Select some $x^0 \in \mathcal{X}$, and $\alpha_0 > 0$
2. Round k of algo $(k \geq 0)$:
   - Output x^k
   - Receive k-th function f_k

Online gradient descent

Algorithm:

1. Select some $x^0 \in \mathcal{X}$, and $\alpha_0 > 0$
2. Round k of algo ($k \geq 0$):
   - Output x^k
   - Receive k-th function f_k
   - Incur loss $f_k(x^k)$

Online gradient descent

Algorithm:

1. Select some $x^0 \in \mathcal{X}$, and $\alpha_0 > 0$
2. Round k of algo ($k \geq 0$):
   - Output x^k
   - Receive k-th function f_k
   - Incur loss $f_k(x^k)$
   - Pick $g^k \in \partial f_k(x_k)$

Online gradient descent

Algorithm:

1. Select some $x^0 \in \mathcal{X}$, and $\alpha_0 > 0$
2. Round k of algo ($k \geq 0$):
   - Output x^k
   - Receive k-th function f_k
   - Incur loss $f_k(x^k)$
   - Pick $g^k \in \partial f_k(x_k)$
     Update: $x^{k+1} = P_{\mathcal{X}}(x^k - \alpha_k g^k)$

Online gradient descent

Algorithm:

1. Select some $x^0 \in \mathcal{X}$, and $\alpha_0 > 0$
2. Round k of algo ($k \geq 0$):
   - Output x^k
   - Receive k-th function f_k
   - Incur loss $f_k(x^k)$
   - Pick $g^k \in \partial f_k(x_k)$

     Update: $x^{k+1} = P_{\mathcal{X}}(x^k - \alpha_k g^k)$

Using $\alpha_k = c/\sqrt{k+1}$ and **assuming** $\|g_k\|_2 \leq G$, can be shown that average regret $\frac{1}{T} R_T \leq O(1/\sqrt{T})$

OGD – regret bound

Assumption: Lipschitz condition $\|\partial f\|_2 \leq G$

OGD – regret bound

Assumption: Lipschitz condition $\|\partial f\|_2 \leq G$

$$x^* = \underset{x \in \mathcal{X}}{\operatorname{argmin}} \sum_{k=1}^{T} f_k(x)$$

OGD – regret bound

Assumption: Lipschitz condition $\|\partial f\|_2 \leq G$

$$x^* = \operatorname*{argmin}_{x \in \mathcal{X}} \sum_{k=1}^{T} f_k(x)$$

Since $g_k \in \partial f_k(x_k)$, we have

$$\boxed{\begin{array}{c} f_k(x^*) \geq f_k(x_k) + \langle g_k,\, x^* - x_k \rangle, \text{ or} \\ f_k(x_k) - f_k(x^*) \leq \langle g_k,\, x_k - x^* \rangle \end{array}}$$

OGD – regret bound

Assumption: Lipschitz condition $\|\partial f\|_2 \le G$

$$x^* = \operatorname*{argmin}_{x \in \mathcal{X}} \sum_{k=1}^{T} f_k(x)$$

Since $g_k \in \partial f_k(x_k)$, we have

$$\boxed{\begin{array}{c} f_k(x^*) \ge f_k(x_k) + \langle g_k,\, x^* - x_k \rangle, \text{ or} \\ f_k(x_k) - f_k(x^*) \le \langle g_k,\, x_k - x^* \rangle \end{array}}$$

Further analysis depends on bounding

$$\boxed{\|x_{k+1} - x^*\|_2^2}$$

OGD regret – bounding distance

Recall: $x_{k+1} = P_{\mathcal{X}}(x_k - \alpha_k g_k)$. Thus,

$$\begin{aligned} \|x_{k+1} - x^*\|_2^2 &= \|P_{\mathcal{X}}(x_k - \alpha_k g_k) - x^*\|_2^2 \\ &= \|P_{\mathcal{X}}(x_k - \alpha_k g_k) - P_{\mathcal{X}}(x^*)\|_2^2 \end{aligned}$$

OGD regret – bounding distance

Recall: $x_{k+1} = P_{\mathcal{X}}(x_k - \alpha_k g_k)$. Thus,

$$\begin{aligned} \|x_{k+1} - x^*\|_2^2 &= \|P_{\mathcal{X}}(x_k - \alpha_k g_k) - x^*\|_2^2 \\ &= \|P_{\mathcal{X}}(x_k - \alpha_k g_k) - P_{\mathcal{X}}(x^*)\|_2^2 \\ (P_{\mathcal{X}} \text{ is nonexpan.}) \quad &\leq \|x_k - x^* - \alpha_k g_k\|_2^2 \end{aligned}$$

OGD regret – bounding distance

Recall: $x_{k+1} = P_{\mathcal{X}}(x_k - \alpha_k g_k)$. Thus,

$$\begin{aligned}
\|x_{k+1} - x^*\|_2^2 &= \|P_{\mathcal{X}}(x_k - \alpha_k g_k) - x^*\|_2^2 \\
&= \|P_{\mathcal{X}}(x_k - \alpha_k g_k) - P_{\mathcal{X}}(x^*)\|_2^2 \\
(P_{\mathcal{X}} \text{ is nonexpan.}) \quad &\leq \|x_k - x^* - \alpha_k g_k\|_2^2 \\
&= \|x_k - x^*\|_2^2 + \alpha_k^2 \|g_k\|_2^2 - 2\alpha_k \langle g_k,\, x_k - x^* \rangle
\end{aligned}$$

OGD regret – bounding distance

Recall: $x_{k+1} = P_{\mathcal{X}}(x_k - \alpha_k g_k)$. Thus,

$$\begin{aligned}
\|x_{k+1} - x^*\|_2^2 &= \|P_{\mathcal{X}}(x_k - \alpha_k g_k) - x^*\|_2^2 \\
&= \|P_{\mathcal{X}}(x_k - \alpha_k g_k) - P_{\mathcal{X}}(x^*)\|_2^2 \\
(P_{\mathcal{X}} \text{ is nonexpan.}) \quad &\leq \|x_k - x^* - \alpha_k g_k\|_2^2 \\
&= \|x_k - x^*\|_2^2 + \alpha_k^2 \|g_k\|_2^2 - 2\alpha_k \langle g_k,\, x_k - x^* \rangle
\end{aligned}$$

$$\langle g_k,\, x_k - x^* \rangle \leq \frac{\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2}{2\alpha_k} + \frac{\alpha_k}{2}\|g_k\|_2^2$$

OGD regret – bounding distance

Recall: $x_{k+1} = P_{\mathcal{X}}(x_k - \alpha_k g_k)$. Thus,

$$\begin{aligned}
\|x_{k+1} - x^*\|_2^2 &= \|P_{\mathcal{X}}(x_k - \alpha_k g_k) - x^*\|_2^2 \\
&= \|P_{\mathcal{X}}(x_k - \alpha_k g_k) - P_{\mathcal{X}}(x^*)\|_2^2 \\
(P_{\mathcal{X}} \text{ is nonexpan.}) \quad &\leq \|x_k - x^* - \alpha_k g_k\|_2^2 \\
&= \|x_k - x^*\|_2^2 + \alpha_k^2 \|g_k\|_2^2 - 2\alpha_k \langle g_k,\, x_k - x^* \rangle
\end{aligned}$$

$$\langle g_k,\, x_k - x^* \rangle \leq \frac{\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2}{2\alpha_k} + \frac{\alpha_k}{2}\|g_k\|_2^2$$

Now invoke $f_k(x_k) - f_k(x^*) \leq \langle g_k,\, x_k - x^* \rangle$

$$f_k(x_k) - f_k(x^*) \leq \frac{\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2}{2\alpha_k} + \frac{\alpha_k}{2}\|g_k\|_2^2$$

OGD regret – bounding distance

Recall: $x_{k+1} = P_{\mathcal{X}}(x_k - \alpha_k g_k)$. Thus,

$$\begin{aligned} \|x_{k+1} - x^*\|_2^2 &= \|P_{\mathcal{X}}(x_k - \alpha_k g_k) - x^*\|_2^2 \\ &= \|P_{\mathcal{X}}(x_k - \alpha_k g_k) - P_{\mathcal{X}}(x^*)\|_2^2 \\ (P_{\mathcal{X}} \text{ is nonexpan.}) \quad &\le \|x_k - x^* - \alpha_k g_k\|_2^2 \\ &= \|x_k - x^*\|_2^2 + \alpha_k^2 \|g_k\|_2^2 - 2\alpha_k \langle g_k,\, x_k - x^* \rangle \end{aligned}$$

$$\langle g_k,\, x_k - x^* \rangle \le \frac{\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2}{2\alpha_k} + \frac{\alpha_k}{2}\|g_k\|_2^2$$

Now invoke $f_k(x_k) - f_k(x^*) \le \langle g_k,\, x_k - x^* \rangle$

$$f_k(x_k) - f_k(x^*) \le \frac{\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2}{2\alpha_k} + \frac{\alpha_k}{2}\|g_k\|_2^2$$

Sum over $k = 1, \ldots, T$, let $\alpha_k = c/\sqrt{k+1}$, use $\|g_k\|_2 \le G$

$$\boxed{\text{Obtain } R_T \le O(\sqrt{T})}$$

References

- ♠ A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. *Robust stochastic approximation approach to stochastic programming*. (2009)
- ♠ J. Linderoth. Lecture slides on *Stochastic Programming* (2003).