Advanced Optimization

(10-801: CMU)

Lecture 21
Incremental methods; Stochastic Optimization

02 Apr 2014

Suvrit Sra

Incremental gradient methods

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

Incremental gradient methods

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

- We saw incremental gradient methods

$$x_{k+1} = x_k - \frac{\eta_k}{m}\nabla f_{i(k)}(x_k), \quad k \geq 0.$$

Incremental gradient methods

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

- We saw incremental gradient methods

$$x_{k+1} = x_k - \frac{\eta_k}{m}\nabla f_{i(k)}(x_k), \quad k \geq 0.$$

- View as gradient-descent with perturbed gradients

$$x_{k+1} = x_k - \frac{\eta_k}{m}(\nabla F(x_k) + e_k)$$

Incremental gradient methods

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

- We saw incremental gradient methods

$$x_{k+1} = x_k - \frac{\eta_k}{m} \nabla f_{i(k)}(x_k), \quad k \geq 0.$$

- View as gradient-descent with perturbed gradients

$$x_{k+1} = x_k - \frac{\eta_k}{m} (\nabla F(x_k) + e_k)$$

- Perturbation slows down rate of convergence. Typically $\eta_k = O(1/k)$; convergence rate also $O(1/k)$ (sublinear).

Incremental gradient methods

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

- We saw incremental gradient methods

$$x_{k+1} = x_k - \frac{\eta_k}{m}\nabla f_{i(k)}(x_k), \quad k \geq 0.$$

- View as gradient-descent with perturbed gradients

$$x_{k+1} = x_k - \frac{\eta_k}{m}(\nabla F(x_k) + e_k)$$

- Perturbation slows down rate of convergence. Typically $\eta_k = O(1/k)$; convergence rate also $O(1/k)$ (sublinear).
- Can we reduce impact of perturbation to speed up?

Stochastic gradients

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

Stochastic gradients

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

The incremental gradient method (IGM)

- Let $x_0 \in \mathbb{R}^n$
- For $k \geq 0$

Stochastic gradients

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

The incremental gradient method (IGM)

- Let $x_0 \in \mathbb{R}^n$
- For $k \geq 0$
  1. Pick $i(k) \in \{1, 2, \ldots, m\}$ uniformly at random
  2. $x_{k+1} = x_k - \eta_k \nabla f_{i(k)}(x_k)$

Stochastic gradients

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

The incremental gradient method (IGM)

- Let $x_0 \in \mathbb{R}^n$
- For $k \geq 0$
  1. Pick $i(k) \in \{1, 2, \ldots, m\}$ uniformly at random
  2. $x_{k+1} = x_k - \eta_k \nabla f_{i(k)}(x_k)$

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

Stochastic gradients

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

The incremental gradient method (IGM)

- Let $x_0 \in \mathbb{R}^n$
- For $k \geq 0$
  1. Pick $i(k) \in \{1, 2, \ldots, m\}$ uniformly at random
  2. $x_{k+1} = x_k - \eta_k \nabla f_{i(k)}(x_k)$

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

$g := g^{\text{true}} + e$, 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:**

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

Stochastic gradients

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

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

Stochastic gradients

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

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

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

Stochastic gradients

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

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

- 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 $\{1, \dots, m\}$
- Thus, **in expectation:**

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

- 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 gradients – more generally

$$x_{k+1} = x_k - \eta_k g_k(x_k, \xi_k),$$

where ξ_k is a rv such that

$$\mathbb{E}_{\xi_k}[g_k(x_k, \xi_k)|x_k] = \nabla F(x_k).$$

Stochastic gradients – more generally

$$x_{k+1} = x_k - \eta_k g_k(x_k, \xi_k),$$

where ξ_k is a rv such that

$$\mathbb{E}_{\xi_k}[g_k(x_k, \xi_k) | x_k] = \nabla F(x_k).$$

- That is, g_k is a **stochastic gradient**.

Stochastic gradients – more generally

$$x_{k+1} = x_k - \eta_k g_k(x_k, \xi_k),$$

where ξ_k is a rv such that

$$\mathbb{E}_{\xi_k}[g_k(x_k, \xi_k) | x_k] = \nabla F(x_k).$$

- That is, g_k is a **stochastic gradient**.

Example: IGM with $g_k = \nabla f_{i(k)}(x_k)$ uses $\xi_k = i(k)$

Stochastic gradients – more generally

$$x_{k+1} = x_k - \eta_k g_k(x_k, \xi_k),$$

where ξ_k is a rv such that

$$\mathbb{E}_{\xi_k}[g_k(x_k, \xi_k) | x_k] = \nabla F(x_k).$$

- That is, g_k is a **stochastic gradient**.

Example: IGM with $g_k = \nabla f_{i(k)}(x_k)$ uses $\xi_k = i(k)$

- g_k equals ∇F only in expectation
- Individual values can vary a lot

Stochastic gradients – more generally

$$x_{k+1} = x_k - \eta_k g_k(x_k, \xi_k),$$

where ξ_k is a rv such that

$$\mathbb{E}_{\xi_k}[g_k(x_k, \xi_k) | x_k] = \nabla F(x_k).$$

- That is, g_k is a **stochastic gradient**.

Example: IGM with $g_k = \nabla f_{i(k)}(x_k)$ uses $\xi_k = i(k)$

- g_k equals ∇F only in expectation
- Individual values can vary a lot
- This variance ($\mathbb{E}[\|g - \nabla F\|^2]$) influences rate of convergence.

Controlling variance

- Instead of using $g_k = \nabla f_{i(k)}(x_k)$, **correct** it by using **true gradient** every m steps (recall: $F = \frac{1}{m}\sum_{i=1}^{m} f_i(x)$)

Controlling variance

- Instead of using $g_k = \nabla f_{i(k)}(x_k)$, **correct** it by using **true gradient** every m steps (recall: $F = \frac{1}{m}\sum_{i=1}^{m} f_i(x)$)
- Reduces variance of $g_k(x_k, \xi_k)$; speeds up convergence

Controlling variance

- Instead of using $g_k = \nabla f_{i(k)}(x_k)$, **correct** it by using **true gradient** every m steps (recall: $F = \frac{1}{m}\sum_{i=1}^m f_i(x)$)
- Reduces variance of $g_k(x_k, \xi_k)$; speeds up convergence

$$\begin{aligned} \nabla F(\bar{x}) &= \frac{1}{m}\sum\nolimits_i f_i(\bar{x}) \\ x_{k+1} &= x_k - \eta_k \underbrace{[\nabla f_{i(k)}(x_k) - \nabla f_{i(k)}(\bar{x}) + \nabla F(\bar{x})]}_{g_k(x_k, \xi_k)} \end{aligned}$$

Controlling variance

- Instead of using $g_k = \nabla f_{i(k)}(x_k)$, **correct** it by using **true gradient** every m steps (recall: $F = \frac{1}{m}\sum_{i=1}^m f_i(x)$)
- Reduces variance of $g_k(x_k, \xi_k)$; speeds up convergence

$$\begin{aligned}\nabla F(\bar{x}) &= \frac{1}{m}\sum_i f_i(\bar{x})\\ x_{k+1} &= x_k - \eta_k \underbrace{[\nabla f_{i(k)}(x_k) - \nabla f_{i(k)}(\bar{x}) + \nabla F(\bar{x})]}_{g_k(x_k,\xi_k)}\end{aligned}$$

- Thus, with $\xi_k = i(k)$, $\mathbb{E}_\xi[g_k|x_k] = \nabla F(x_k)$

Controlling variance

- Instead of using $g_k = \nabla f_{i(k)}(x_k)$, **correct** it by using **true gradient** every m steps (recall: $F = \frac{1}{m}\sum_{i=1}^{m} f_i(x)$)
- Reduces variance of $g_k(x_k, \xi_k)$; speeds up convergence

$$\begin{aligned} \nabla F(\bar{x}) &= \frac{1}{m}\sum_i f_i(\bar{x}) \\ x_{k+1} &= x_k - \eta_k \underbrace{[\nabla f_{i(k)}(x_k) - \nabla f_{i(k)}(\bar{x}) + \nabla F(\bar{x})]}_{g_k(x_k, \xi_k)} \end{aligned}$$

- Thus, with $\xi_k = i(k)$, $\mathbb{E}_\xi[g_k | x_k] = \nabla F(x_k)$

Same expectation, lower variance

Controlling variance

- Instead of using $g_k = \nabla f_{i(k)}(x_k)$, **correct** it by using **true gradient** every m steps (recall: $F = \frac{1}{m}\sum_{i=1}^{m} f_i(x)$)
- Reduces variance of $g_k(x_k, \xi_k)$; speeds up convergence

$$\begin{aligned} \nabla F(\bar{x}) &= \frac{1}{m}\sum_i f_i(\bar{x}) \\ x_{k+1} &= x_k - \eta_k \underbrace{[\nabla f_{i(k)}(x_k) - \nabla f_{i(k)}(\bar{x}) + \nabla F(\bar{x})]}_{g_k(x_k,\xi_k)} \end{aligned}$$

- Thus, with $\xi_k = i(k)$, $\mathbb{E}_\xi[g_k|x_k] = \nabla F(x_k)$

Same expectation, lower variance

Say $\bar{x}, x_k \to x^*$. Then $\nabla F(\bar{x}) \to 0$.

Controlling variance

- Instead of using $g_k = \nabla f_{i(k)}(x_k)$, **correct** it by using **true gradient** every m steps (recall: $F = \frac{1}{m}\sum_{i=1}^m f_i(x)$)
- Reduces variance of $g_k(x_k, \xi_k)$; speeds up convergence

$$\begin{aligned} \nabla F(\bar{x}) &= \tfrac{1}{m}\sum\nolimits_i f_i(\bar{x}) \\ x_{k+1} &= x_k - \eta_k \underbrace{[\nabla f_{i(k)}(x_k) - \nabla f_{i(k)}(\bar{x}) + \nabla F(\bar{x})]}_{g_k(x_k,\xi_k)} \end{aligned}$$

- Thus, with $\xi_k = i(k)$, $\mathbb{E}_\xi[g_k|x_k] = \nabla F(x_k)$

Same expectation, lower variance

Say $\bar{x}, x_k \to x^*$. Then $\nabla F(\bar{x}) \to 0$. Thus, if $\nabla f_i(\bar{x}) \to \nabla f_i(x^*)$, then

Controlling variance

- Instead of using $g_k = \nabla f_{i(k)}(x_k)$, **correct** it by using **true gradient** every m steps (recall: $F = \frac{1}{m}\sum_{i=1}^m f_i(x)$)
- Reduces variance of $g_k(x_k, \xi_k)$; speeds up convergence

$$\begin{aligned}\nabla F(\bar{x}) &= \frac{1}{m}\sum_i f_i(\bar{x})\\ x_{k+1} &= x_k - \eta_k \underbrace{[\nabla f_{i(k)}(x_k) - \nabla f_{i(k)}(\bar{x}) + \nabla F(\bar{x})]}_{g_k(x_k,\xi_k)}\end{aligned}$$

- Thus, with $\xi_k = i(k)$, $\mathbb{E}_\xi[g_k|x_k] = \nabla F(x_k)$

Same expectation, lower variance

Say $\bar{x}, x_k \to x^*$. Then $\nabla F(\bar{x}) \to 0$. Thus, if $\nabla f_i(\bar{x}) \to \nabla f_i(x^*)$, then

$$\nabla f_i(x_k) - \nabla f_i(\bar{x}) + \nabla F(\bar{x}) \to \nabla f_i(x_k) - \nabla f_i(x^*) \to 0.$$

SG with variance reduction

- For $s \geq 1$:
  1. $\bar{x} \leftarrow \bar{x}_{s-1}$
  2. $\bar{g} \leftarrow \nabla F(\bar{x})$ (full gradient computation)

SG with variance reduction

- For $s \geq 1$:
  1. $\bar{x} \leftarrow \bar{x}_{s-1}$
  2. $\bar{g} \leftarrow \nabla F(\bar{x})$ (full gradient computation)
  3. $x_0 = \bar{x}; \quad t \leftarrow \textsc{Rand}(1, m)$ (randomized stopping)

SG with variance reduction

- For $s \geq 1$:
  1. $\bar{x} \leftarrow \bar{x}_{s-1}$
  2. $\bar{g} \leftarrow \nabla F(\bar{x})$ (full gradient computation)
  3. $x_0 = \bar{x}; \quad t \leftarrow \text{RAND}(1, m)$ (randomized stopping)
  4. For $k = 0, 1, \ldots, t-1$
     - Randomly pick $i(k) \in [1..m]$
     - $x_{k+1} = x_k - \eta_k(\nabla f_{i(k)}(x_k) - \nabla f_{i(k)}(\bar{x}) + \bar{g})$

SG with variance reduction

- For $s \geq 1$:
  1. $\bar{x} \leftarrow \bar{x}_{s-1}$
  2. $\bar{g} \leftarrow \nabla F(\bar{x})$ (full gradient computation)
  3. $x_0 = \bar{x}; \quad t \leftarrow \text{RAND}(1, m)$ (randomized stopping)
  4. For $k = 0, 1, \ldots, t-1$
     - Randomly pick $i(k) \in [1..m]$
     - $x_{k+1} = x_k - \eta_k(\nabla f_{i(k)}(x_k) - \nabla f_{i(k)}(\bar{x}) + \bar{g})$
  5. $\bar{x}_s \leftarrow x_t$

SG with variance reduction

- For $s \geq 1$:
  1. $\bar{x} \leftarrow \bar{x}_{s-1}$
  2. $\bar{g} \leftarrow \nabla F(\bar{x})$ (full gradient computation)
  3. $x_0 = \bar{x}; \quad t \leftarrow \text{RAND}(1, m)$ (randomized stopping)
  4. For $k = 0, 1, \ldots, t-1$
     - Randomly pick $i(k) \in [1..m]$
     - $x_{k+1} = x_k - \eta_k(\nabla f_{i(k)}(x_k) - \nabla f_{i(k)}(\bar{x}) + \bar{g})$
  5. $\bar{x}_s \leftarrow x_t$

Theorem Assume each $f_i(x)$ is smooth convex and $F(x)$ is strongly-convex. Then, for sufficiently large n, there is $\alpha < 1$ s.t.

$$\mathbb{E}[F(\bar{x}_s) - F(x^*)] \leq \alpha^s [F(\bar{x}_0) - F(x^*)]$$

SG with variance reduction

- For $s \geq 1$:
  1. $\bar{x} \leftarrow \bar{x}_{s-1}$
  2. $\bar{g} \leftarrow \nabla F(\bar{x})$ (full gradient computation)
  3. $x_0 = \bar{x}; \quad t \leftarrow \text{RAND}(1, m)$ (randomized stopping)
  4. For $k = 0, 1, \ldots, t-1$
     - Randomly pick $i(k) \in [1..m]$
     - $x_{k+1} = x_k - \eta_k(\nabla f_{i(k)}(x_k) - \nabla f_{i(k)}(\bar{x}) + \bar{g})$
  5. $\bar{x}_s \leftarrow x_t$

Theorem Assume each $f_i(x)$ is smooth convex and $F(x)$ is strongly-convex. Then, for sufficiently large n, there is $\alpha < 1$ s.t.

$$\mathbb{E}[F(\bar{x}_s) - F(x^*)] \leq \alpha^s [F(\bar{x}_0) - F(x^*)]$$

Rmk: Typically for stochastic methods we make stmts of the form

$$\mathbb{E}[F(x_k) - F(x^*)] \leq O(1/k)$$

Stochastic Optimization

Stochastic optimization – example

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 $\omega_2 \sim \mathcal{U}[1/3,1]$

Stochastic optimization – example

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

Stochastic optimization – example

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?

Stochastic optimization – example

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 optimization – example

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 optimization – example

- But we cannot “wait-and-watch” —

Stochastic optimization – example

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

Stochastic optimization – example

- 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 optimization – example

- 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 optimization – 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
- *None of these:* **Sample...**

Stochastic optimization – 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 $\omega_2 \sim \mathcal{U}[1/3,1]$

Unbiased / Average case:

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

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

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

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

Stochastic optimization – 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 $\omega_2 \sim \mathcal{U}[1/3,1]$

Worst case:

$$\omega_1 = 1, \quad \omega_2 = 1/3$$

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

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

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

Stochastic optimization – 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:

$$\omega_1 = 5, \quad \mathbb{E}[\omega_2] = 1$$

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

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

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

Stochastic optimization via sampling

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

- ξ follows some **known** distribution

Stochastic optimization via sampling

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

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

Stochastic optimization via sampling

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

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

Stochastic optimization via sampling

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

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

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

is a difficult high-dimensional integral.

Stochastic optimization – setup

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

Setup and Assumptions

1. $\mathcal{X} \subset \mathbb{R}^n$ compact convex set

Stochastic optimization – setup

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

Setup and Assumptions

1. $\mathcal{X} \subset \mathbb{R}^n$ compact convex set

2. ξ 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 optimization – setup

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

Setup and Assumptions

1. $\mathcal{X} \subset \mathbb{R}^n$ compact convex set

2. ξ 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

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

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

Stochastic optimization – setup

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

Setup and Assumptions

1. $\mathcal{X} \subset \mathbb{R}^n$ compact convex set

2. ξ 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

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

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

4. For every $\xi \in \Omega$, $f(\cdot, \xi)$ is convex.

Convex stochastic optimization problem

Stochastic optimization – setup

- Cannot compute expectation in general

Stochastic optimization – setup

- Cannot compute expectation in general
- Computational techniques based on sampling

Stochastic optimization – setup

- Cannot compute expectation in general
- Computational techniques based on sampling

Assumption 1: Possible to generate independent identically distributed (iid) samples $\xi_1, \xi_2, \ldots$
Assumption 2: For pair $(x, \xi) \in \mathcal{X} \times \Omega$, oracle yields stochastic gradient $g(x, \xi)$, i.e.,

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

Stochastic optimization – setup

- Cannot compute expectation in general
- Computational techniques based on sampling

Assumption 1: Possible to generate independent identically distributed (iid) samples $\xi_1, \xi_2, \ldots$
Assumption 2: For pair $(x, \xi) \in \mathcal{X} \times \Omega$, oracle yields stochastic gradient $g(x, \xi)$, i.e.,

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

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

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

Stochastic optimization – setup

- Cannot compute expectation in general
- Computational techniques based on sampling

Assumption 1: Possible to generate independent identically distributed (iid) samples $\xi_1, \xi_2, \dots$
Assumption 2: For pair $(x, \xi) \in \mathcal{X} \times \Omega$, oracle yields stochastic gradient $g(x, \xi)$, i.e.,

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

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

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

- So $g(x, \omega) \in \partial_x f(x, \omega)$ is a stochastic subgradient.

Stochastic optimization – approaches

- ♣ Stochastic Approximation (SA)
  - Sample ξ_k iid

Stochastic optimization – approaches

- ♣ Stochastic Approximation (SA)
  - Sample ξ_k iid
  - Generate stochastic subgradient $g(x, \xi)$

Stochastic optimization – approaches

- ♣ Stochastic Approximation (SA)
  - Sample ξ_k iid
  - Generate stochastic subgradient $g(x, \xi)$
  - Use that in a subgradient method

Stochastic optimization – approaches

- ♣ Stochastic Approximation (SA)
  - ▶ Sample ξ_k iid
  - ▶ Generate stochastic subgradient $g(x, \xi)$
  - ▶ Use that in a subgradient method
- ♣ Sample average approximation (SAA)

Stochastic optimization – approaches

- ♣ Stochastic Approximation (SA)
  - Sample ξ_k iid
  - Generate stochastic subgradient $g(x, \xi)$
  - Use that in a subgradient method
- ♣ Sample average approximation (SAA)
  - Generate m iid samples, $\xi_1, \ldots, \xi_m$

Stochastic optimization – approaches

♣ Stochastic Approximation (SA)

- Sample ξ_k iid
- Generate stochastic subgradient $g(x, \xi)$
- Use that in a subgradient method

♣ Sample average approximation (SAA)

- Generate m iid samples, $\xi_1, \ldots, \xi_m$
- Consider **empirical objective** $\hat{F}_m := m^{-1} \sum_i f(x, \xi_i)$

Stochastic optimization – approaches

- Stochastic Approximation (SA)
  - Sample ξ_k iid
  - Generate stochastic subgradient $g(x, \xi)$
  - Use that in a subgradient method
- Sample average approximation (SAA)
  - Generate m iid samples, $\xi_1, \ldots, \xi_m$
  - Consider **empirical objective** $\hat{F}_m := m^{-1} \sum_i f(x, \xi_i)$
  - SAA refers to creation of this **sample average problem**
  - Minimizing $\hat{F}_m$ 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; obtain $g(x_k, \xi_k)$ from oracle
  - Update $x_{k+1} = P_{\mathcal{X}}(x_k - \alpha_k g(x_k, \xi_k))$, 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; obtain $g(x_k, \xi_k)$ from oracle
  - Update $x_{k+1} = P_{\mathcal{X}}(x_k - \alpha_k g(x_k, \xi_k))$, where $\alpha_k > 0$

We'll simply write

$$\boxed{x_{k+1} = P_{\mathcal{X}}\big(x_k - \alpha_k g_k\big)}$$

Stochastic approximation – SA

SA or stochastic (sub)-gradient

- Let $x_0 \in \mathcal{X}$
- For $k \geq 0$
  - Sample ω_k; obtain $g(x_k, \xi_k)$ from oracle
  - Update $x_{k+1} = P_{\mathcal{X}}(x_k - \alpha_k g(x_k, \xi_k))$, where $\alpha_k > 0$

We'll simply write

$$x_{k+1} = P_{\mathcal{X}}\big(x_k - \alpha_k g_k\big)$$

Does this work?

Stochastic approximation – analysis

Setup

- x_k depends on rvs $\xi_1, \ldots, \xi_{k-1}$, so itself random

Stochastic approximation – analysis

Setup

- x_k depends on rvs $\xi_1, \ldots, \xi_{k-1}$, so itself random
- Of course, x_k does not depend on ξ_k

Stochastic approximation – analysis

Setup

- x_k depends on rvs $\xi_1, \ldots, \xi_{k-1}$, so itself random
- Of course, x_k does not depend on ξ_k
- Subgradient method analysis hinges upon: $\|x_k - x^*\|^2$

Stochastic approximation – analysis

Setup

- x_k depends on rvs $\xi_1, \ldots, \xi_{k-1}$, so itself random
- Of course, x_k does not depend on ξ_k
- Subgradient method analysis hinges upon: $\|x_k - x^*\|^2$
- Stochastic subgradient hinges upon: $\mathbb{E}[\|x_k - x^*\|^2]$

Stochastic approximation – analysis

Setup

- x_k depends on rvs $\xi_1, \ldots, \xi_{k-1}$, so itself random
- Of course, x_k does not depend on ξ_k
- Subgradient method analysis hinges upon: $\|x_k - x^*\|^2$
- Stochastic subgradient hinges upon: $\mathbb{E}[\|x_k - x^*\|^2]$

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

Stochastic approximation – analysis

Setup

- x_k depends on rvs $\xi_1, \ldots, \xi_{k-1}$, so itself random
- Of course, x_k does not depend on ξ_k
- Subgradient method analysis hinges upon: $\|x_k - x^*\|^2$
- Stochastic subgradient hinges upon: $\mathbb{E}[\|x_k - x^*\|^2]$

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

Bounding R_{k+1}

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

Stochastic approximation – analysis

Setup

- x_k depends on rvs $\xi_1, \ldots, \xi_{k-1}$, so itself random
- Of course, x_k does not depend on ξ_k
- Subgradient method analysis hinges upon: $\|x_k - x^*\|^2$
- Stochastic subgradient hinges upon: $\mathbb{E}[\|x_k - x^*\|^2]$

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

Bounding R_{k+1}

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

Stochastic approximation – analysis

Setup

- x_k depends on rvs $\xi_1, \ldots, \xi_{k-1}$, so itself random
- Of course, x_k does not depend on ξ_k
- Subgradient method analysis hinges upon: $\|x_k - x^*\|^2$
- Stochastic subgradient hinges upon: $\mathbb{E}[\|x_k - x^*\|^2]$

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

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

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

- **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

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

- **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

Stochastic approximation – analysis

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

- **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 ξ_k, we have

$$\mathbb{E}[\langle x_k - x^*,\, g(x_k, \xi_k) \rangle] \quad =$$

Stochastic approximation – analysis

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

- **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 ξ_k, we have

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

Stochastic approximation – analysis

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

- **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 ξ_k, we have

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

Stochastic approximation – analysis

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

- **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 ξ_k, we have

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

Stochastic approximation – analysis

It remains to bound: $\mathbb{E}[\langle x_k - x^*, G_k \rangle]$

Stochastic approximation – analysis

It remains 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

It remains 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

$$2\alpha_k \mathbb{E}[F(x^*) - F(x_k)] \geq 2\alpha_k \mathbb{E}[\langle G_k, x^* - x_k \rangle]$$

Stochastic approximation – analysis

It remains 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

$$2\alpha_k \mathbb{E}[F(x^*) - F(x_k)] \geq 2\alpha_k \mathbb{E}[\langle G_k, x^* - x_k \rangle]$$

Plug this bound back into the r_{k+1} inequality:

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

Stochastic approximation – analysis

It remains 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

$$2\alpha_k \mathbb{E}[F(x^*) - F(x_k)] \geq 2\alpha_k \mathbb{E}[\langle G_k, x^* - x_k\rangle]$$

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

It remains 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

$$2\alpha_k \mathbb{E}[F(x^*) - F(x_k)] \geq 2\alpha_k \mathbb{E}[\langle G_k, x^* - x_k\rangle]$$

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

It remains 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

$$2\alpha_k \mathbb{E}[F(x^*) - F(x_k)] \geq 2\alpha_k \mathbb{E}[\langle G_k, x^* - x_k\rangle]$$

Plug this bound back into the r_{k+1} inequality:

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

We've bounded the expected progress; 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 $i = 1, \ldots, k$, to obtain

$$\sum_{i=1}^{k} (2\alpha_i \mathbb{E}[F(x_i) - f(x^*)]) \quad \leq \quad r_1 - r_{k+1} + M^2 \sum_i \alpha_i^2$$

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 $i = 1, \ldots, k$, to obtain

$$\begin{aligned} \sum_{i=1}^{k} (2\alpha_i \mathbb{E}[F(x_i) - f(x^*)]) &\le r_1 - r_{k+1} + M^2 \sum_i \alpha_i^2 \\ &\le r_1 + M^2 \sum_i \alpha_i^2. \end{aligned}$$

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 $i = 1, \ldots, k$, to obtain

$$\begin{aligned} \sum_{i=1}^{k} (2\alpha_i \mathbb{E}[F(x_i) - f(x^*)]) &\leq r_1 - r_{k+1} + M^2 \sum_i \alpha_i^2 \\ &\leq r_1 + M^2 \sum_i \alpha_i^2. \end{aligned}$$

Divide both sides by $\sum_i \alpha_i$, so

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 $i = 1, \ldots, k$, to obtain

$$\begin{aligned}\sum\nolimits_{i=1}^{k}(2\alpha_i\mathbb{E}[F(x_i) - f(x^*)]) &\le r_1 - r_{k+1} + M^2\sum\nolimits_i \alpha_i^2\\ &\le r_1 + M^2\sum\nolimits_i \alpha_i^2.\end{aligned}$$

Divide both sides by $\sum_i \alpha_i$, so

- Set $\gamma_i = \frac{\alpha_i}{\sum_i^k \alpha_i}$.
- Thus, $\gamma_i \ge 0$ and $\sum_i \gamma_i = 1$

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 $i = 1, \ldots, k$, to obtain

$$\begin{aligned} \sum_{i=1}^{k} (2\alpha_i \mathbb{E}[F(x_i) - f(x^*)]) &\leq r_1 - r_{k+1} + M^2 \sum_i \alpha_i^2 \\ &\leq r_1 + M^2 \sum_i \alpha_i^2. \end{aligned}$$

Divide both sides by $\sum_i \alpha_i$, so

- Set $\gamma_i = \frac{\alpha_i}{\sum_i^k \alpha_i}$.
- Thus, $\gamma_i \geq 0$ and $\sum_i \gamma_i = 1$

$$\mathbb{E}\left[\sum_i \gamma_i (F(x_i) - F(x^*))\right] \leq \frac{r_1 + M^2 \sum_i \alpha_i^2}{2 \sum_i \alpha_i}$$

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_k

Stochastic approximation – analysis

- Bound looks similar to bound in subgradient method
- But we wish to say something about x_k
- Since $\gamma_i \geq 0$ and $\sum_i^k \gamma_i = 1$, and we have $\gamma_i F(x_i)$

Stochastic approximation – analysis

- Bound looks similar to bound in subgradient method
- But we wish to say something about x_k
- Since $\gamma_i \geq 0$ and $\sum_i^k \gamma_i = 1$, and we have $\gamma_i F(x_i)$
- Easier to talk about **averaged**

$$\bar{x}_k := \sum_i^k \gamma_i x_i.$$

Stochastic approximation – analysis

- Bound looks similar to bound in subgradient method
- But we wish to say something about x_k
- Since $\gamma_i \geq 0$ and $\sum_i^k \gamma_i = 1$, and we have $\gamma_i F(x_i)$
- Easier to talk about **averaged**

$$\bar{x}_k := \sum\nolimits_i^k \gamma_i x_i.$$

- $f(\bar{x}_k) \leq \sum_i \gamma_i F(x_i)$ due to convexity

Stochastic approximation – analysis

- Bound looks similar to bound in subgradient method
- But we wish to say something about x_k
- Since $\gamma_i \geq 0$ and $\sum_i^k \gamma_i = 1$, and we have $\gamma_i F(x_i)$
- Easier to talk about **averaged**

$$\bar{x}_k := \sum_i^k \gamma_i x_i.$$

- $f(\bar{x}_k) \leq \sum_i \gamma_i F(x_i)$ due to convexity
- So we finally obtain the inequality

$$\mathbb{E}\big[F(\bar{x}_k) - F(x^*)\big] \leq \frac{r_1 + M^2 \sum_i \alpha_i^2}{2 \sum_i \alpha_i}.$$

Stochastic approximation – finally

- Let $D_{\mathcal{X}} := \max_{x \in \mathcal{X}} \|x - x^*\|_2$ (act. only need $\|x_1 - x^*\| \leq D_{\mathcal{X}}$)
- Assume $\alpha_i = \alpha$ is a constant. Observe that

$$\mathbb{E}[F(\bar{x}_k) - F(x^*)] \leq \frac{D_{\mathcal{X}}^2 + M^2 k \alpha^2}{2k\alpha}$$

- Minimize the rhs over $\alpha > 0$ to obtain $\mathbb{E}[F(\bar{x}_k) - F(x^*)] \leq \frac{D_{\mathcal{X}} M}{\sqrt{k}}$
- If k is not fixed in advance, then choose

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

- Analyze $\mathbb{E}[F(\bar{x}_k) - F(x^*)]$ with this choice of stepsize

Stochastic approximation – finally

- Let $D_{\mathcal{X}} := \max_{x \in \mathcal{X}} \|x - x^*\|_2$ (act. only need $\|x_1 - x^*\| \leq D_{\mathcal{X}}$)
- Assume $\alpha_i = \alpha$ is a constant. Observe that

$$\mathbb{E}[F(\bar{x}_k) - F(x^*)] \leq \frac{D_{\mathcal{X}}^2 + M^2 k \alpha^2}{2k\alpha}$$

- Minimize the rhs over $\alpha > 0$ to obtain $\mathbb{E}[F(\bar{x}_k) - F(x^*)] \leq \frac{D_{\mathcal{X}} M}{\sqrt{k}}$
- If k is not fixed in advance, then choose

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

- Analyze $\mathbb{E}[F(\bar{x}_k) - F(x^*)]$ with this choice of stepsize

We showed $O(1/\sqrt{k})$ rate

Stochastic approximation – remarks

Theorem Let $f(x,\xi)$ be C^1_L convex. Let $e_k := \nabla F(x_k) - g_k$ satisfy $\mathbb{E}[e_k] = 0$. Let $\|x_i - x^*\| \le D$. Also, let $\alpha_i = 1/(L+\eta_i)$. Then,

$$\mathbb{E}\Big[\sum_{i=1}^{k} F(x_{i+1}) - F(x^*)\Big] \le \tfrac{D^2}{2\alpha_k} + \sum_{i=1}^{k} \tfrac{\mathbb{E}[\|e_i\|^2]}{2\eta_i}.$$

Stochastic approximation – remarks

Theorem Let $f(x,\xi)$ be C_L^1 convex. Let $e_k := \nabla F(x_k) - g_k$ satisfy $\mathbb{E}[e_k] = 0$. Let $\|x_i - x^*\| \le D$. Also, let $\alpha_i = 1/(L+\eta_i)$. Then,

$$\mathbb{E}\left[\sum_{i=1}^{k} F(x_{i+1}) - F(x^*)\right] \le \frac{D^2}{2\alpha_k} + \sum_{i=1}^{k} \frac{\mathbb{E}[\|e_i\|^2]}{2\eta_i}.$$

As before, by using $\bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_{i+1}$ we get

$$\mathbb{E}[F(\bar{x}_k) - F(x^*)] \le \frac{D^2}{2\alpha_k k} + \frac{1}{k}\sum_{i=1}^{k} \frac{\mathbb{E}[\|e_i\|^2]}{2\eta_i}.$$

Stochastic approximation – remarks

Theorem Let $f(x,\xi)$ be C^1_L convex. Let $e_k := \nabla F(x_k) - g_k$ satisfy $\mathbb{E}[e_k] = 0$. Let $\|x_i - x^*\| \leq D$. Also, let $\alpha_i = 1/(L+\eta_i)$. Then,

$$\mathbb{E}\Big[\sum_{i=1}^{k} F(x_{i+1}) - F(x^*)\Big] \leq \frac{D^2}{2\alpha_k} + \sum_{i=1}^{k} \frac{\mathbb{E}[\|e_i\|^2]}{2\eta_i}.$$

As before, by using $\bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_{i+1}$ we get

$$\mathbb{E}[F(\bar{x}_k) - F(x^*)] \leq \frac{D^2}{2\alpha_k k} + \frac{1}{k}\sum_{i=1}^{k} \frac{\mathbb{E}[\|e_i\|^2]}{2\eta_i}.$$

- Using $\alpha_i = L + \eta_i$ where $\eta_i \propto 1/\sqrt{i}$ we obtain

Stochastic approximation – remarks

Theorem Let $f(x,\xi)$ be C^1_L convex. Let $e_k := \nabla F(x_k) - g_k$ satisfy $\mathbb{E}[e_k] = 0$. Let $\|x_i - x^*\| \leq D$. Also, let $\alpha_i = 1/(L+\eta_i)$. Then,

$$\mathbb{E}\big[\sum\nolimits_{i=1}^{k} F(x_{i+1}) - F(x^*)\big] \leq \frac{D^2}{2\alpha_k} + \sum\nolimits_{i=1}^{k} \frac{\mathbb{E}[\|e_i\|^2]}{2\eta_i}.$$

As before, by using $\bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_{i+1}$ we get

$$\mathbb{E}[F(\bar{x}_k) - F(x^*)] \leq \frac{D^2}{2\alpha_k k} + \frac{1}{k}\sum\nolimits_{i=1}^{k} \frac{\mathbb{E}[\|e_i\|^2]}{2\eta_i}.$$

- Using $\alpha_i = L + \eta_i$ where $\eta_i \propto 1/\sqrt{i}$ we obtain

$$\mathbb{E}[F(\bar{x}_k) - F(x^*)] = O(\tfrac{LD^2}{k}) + O(\tfrac{\sigma D}{\sqrt{k}})$$

where σ bounds the variance $\mathbb{E}[\|e_i\|^2]$

Stochastic approximation – remarks

Theorem Let $f(x,\xi)$ be C^1_L convex. Let $e_k := \nabla F(x_k) - g_k$ satisfy $\mathbb{E}[e_k] = 0$. Let $\|x_i - x^*\| \leq D$. Also, let $\alpha_i = 1/(L + \eta_i)$. Then,

$$\mathbb{E}\Big[\sum_{i=1}^{k} F(x_{i+1}) - F(x^*)\Big] \leq \frac{D^2}{2\alpha_k} + \sum_{i=1}^{k} \frac{\mathbb{E}[\|e_i\|^2]}{2\eta_i}.$$

As before, by using $\bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_{i+1}$ we get

$$\mathbb{E}[F(\bar{x}_k) - F(x^*)] \leq \frac{D^2}{2\alpha_k k} + \frac{1}{k}\sum_{i=1}^{k} \frac{\mathbb{E}[\|e_i\|^2]}{2\eta_i}.$$

- Using $\alpha_i = L + \eta_i$ where $\eta_i \propto 1/\sqrt{i}$ we obtain

$$\mathbb{E}[F(\bar{x}_k) - F(x^*)] = O\Big(\frac{LD^2}{k}\Big) + O\Big(\frac{\sigma D}{\sqrt{k}}\Big)$$

where σ bounds the variance $\mathbb{E}[\|e_i\|^2]$

Minimax optimal rate

Stochastic approximation – remarks

Theorem Suppose $f(x,\xi)$ are convex and $F(x)$ is μ-strongly convex. Let $\bar{x}_k := \sum_{i=0}^{k} \theta_i x_i$, where $\theta_i = \frac{2(i+1)}{(k+1)(k+2)}$, we obtain

$$\mathbb{E}[F(\bar{x}_k) - F(x^*)] \leq \frac{2M^2}{\mu^2(k+1)}.$$

Lacoste-Julien, Schmidt, Bach (2012).

Stochastic approximation – remarks

Theorem Suppose $f(x,\xi)$ are convex and $F(x)$ is μ-strongly convex. Let $\bar{x}_k := \sum_{i=0}^{k} \theta_i x_i$, where $\theta_i = \frac{2(i+1)}{(k+1)(k+2)}$, we obtain

$$\mathbb{E}[F(\bar{x}_k) - F(x^*)] \leq \frac{2M^2}{\mu^2(k+1)}.$$

Lacoste-Julien, Schmidt, Bach (2012).
With uniform averaging $\bar{x}_k = \frac{1}{k}\sum_i x_i$, we get $O(\log k/k)$.

Sample average approximation

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

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

Sample Average Approximation (SAA):

- Collect samples $\xi_1, \ldots, \omega_m$
- Empirical objective: $\hat{F}_m(x) := \frac{1}{m}\sum_{i=1}^m f(x,\xi_i)$

Sample average approximation

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

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

Sample Average Approximation (SAA):

- Collect samples $\xi_1, \ldots, \omega_m$
- Empirical objective: $\hat{F}_m(x) := \frac{1}{m} \sum_{i=1}^{m} f(x, \xi_i)$
- aka *Empirical Risk Minimization*

Sample average approximation

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

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

Sample Average Approximation (SAA):

- Collect samples $\xi_1, \ldots, \omega_m$
- Empirical objective: $\hat{F}_m(x) := \frac{1}{m}\sum_{i=1}^{m} f(x,\xi_i)$
- aka *Empirical Risk Minimization*
- Confusing: We often optimize $\hat{F}_m$ using stochastic subgradient; but theoretical guarantees are then only on the *empirical* suboptimality $E[\hat{F}_m(\bar{x}_k)] \le \ldots$

Sample average approximation

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

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

Sample Average Approximation (SAA):

- Collect samples $\xi_1, \ldots, \omega_m$
- Empirical objective: $\hat{F}_m(x) := \frac{1}{m}\sum_{i=1}^m f(x,\xi_i)$
- aka *Empirical Risk Minimization*
- Confusing: We often optimize $\hat{F}_m$ using stochastic subgradient; but theoretical guarantees are then only on the *empirical* suboptimality $E[\hat{F}_m(\bar{x}_k)] \le \ldots$
- For guarantees on $F(\bar{x}_k)$ more work; (*regularization* + conc.) $F(\bar{x}_k) - F(x^*) \le O(1/\sqrt{k}) + O(1/\sqrt{m})$

Online optimization

Online optimization

- We have *fixed* and *known* $f(x, \xi)$

Online optimization

- We have *fixed* and *known* $f(x, \xi)$
- $\xi_1, \xi_2, \ldots$ presented to us sequentially

Can be chosen adversarially!

Online optimization

- We have *fixed* and *known* $f(x, \xi)$
- $\xi_1, \xi_2, \ldots$ presented to us sequentially

Can be chosen adversarially!

- **Guess** x_k;

Online optimization

- We have *fixed* and *known* $f(x, \xi)$
- $\xi_1, \xi_2, \ldots$ presented to us sequentially

Can be chosen adversarially!

- **Guess** x_k; **Observe** ξ_k;

Online optimization

- We have *fixed* and *known* $f(x, \xi)$
- $\xi_1, \xi_2, \ldots$ presented to us sequentially

Can be chosen adversarially!

- **Guess** x_k; **Observe** ξ_k; **incur cost** $f(x_k, \xi_k)$;

Online optimization

- We have *fixed* and *known* $f(x, \xi)$
- $\xi_1, \xi_2, \ldots$ presented to us sequentially

  Can be chosen adversarially!

- **Guess** x_k; **Observe** ξ_k; **incur cost** $f(x_k, \xi_k)$; **Update** to x_{k+1}

Online optimization

- We have *fixed* and *known* $f(x, \xi)$
- $\xi_1, \xi_2, \ldots$ presented to us sequentially

  Can be chosen adversarially!

- **Guess** x_k; **Observe** ξ_k; **incur cost** $f(x_k, \xi_k)$; **Update** to x_{k+1}
- We get to see things only sequentially; sequence of samples shown to us by nature may depend on our guesses

Online optimization

- We have *fixed* and *known* $f(x, \xi)$
- $\xi_1, \xi_2, \ldots$ presented to us sequentially

Can be chosen adversarially!

- **Guess** x_k; **Observe** ξ_k; **incur cost** $f(x_k, \xi_k)$; **Update** to x_{k+1}
- We get to see things only sequentially; 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* $f(x, \xi)$
- $\xi_1, \xi_2, \ldots$ presented to us sequentially

Can be chosen adversarially!

- **Guess** x_k; **Observe** ξ_k; **incur cost** $f(x_k, \xi_k)$; **Update** to x_{k+1}
- We get to see things only sequentially; 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* $f(x, \xi)$
- $\xi_1, \xi_2, \ldots$ presented to us sequentially

Can be chosen adversarially!

- **Guess** x_k; **Observe** ξ_k; **incur cost** $f(x_k, \xi_k)$; **Update** to x_{k+1}
- We get to see things only sequentially; 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* $f(x, \xi)$
- $\xi_1, \xi_2, \ldots$ presented to us sequentially

Can be chosen adversarially!

- **Guess** x_k; **Observe** ξ_k; **incur cost** $f(x_k, \xi_k)$; **Update** to x_{k+1}
- We get to see things only sequentially; 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 is an important idea in machine learning, game theory, decision making, etc.

Online gradient descent

Based on Zinkevich (2003)

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

Online gradient descent

Based on Zinkevich (2003)

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

Simplify notation: $f_k(x) \equiv f(x, \xi_k)$

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 $\|\partial f\|_2 \leq G$

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

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

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

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

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

Further analysis depends on bounding

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

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,

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

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

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

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

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

Obtain $R_T \leq O(\sqrt{T})$

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).