Outline - I

1. **Introduction**
   - Large-scale machine learning and optimization
   - Classes of functions (convex, smooth, etc.)
   - Traditional statistical analysis through Rademacher complexity

2. **Classical methods for convex optimization**
   - Smooth optimization (gradient descent, Newton method)
   - Non-smooth optimization (subgradient descent)
   - Proximal methods

3. **Non-smooth stochastic approximation**
   - Stochastic (sub)gradient and averaging
   - Non-asymptotic results and lower bounds
   - Strongly convex vs. non-strongly convex

Outline - II

4. **Classical stochastic approximation**
   - Asymptotic analysis
   - Robbins-Monro algorithm
   - Polyak-Rupert averaging

5. **Smooth stochastic approximation algorithms**
   - Non-asymptotic analysis for smooth functions
   - Logistic regression
   - Least-squares regression without decaying step-sizes

6. **Finite data sets**
   - Gradient methods with exponential convergence rates
   - Convex duality
   - (Dual) stochastic coordinate descent - Frank-Wolfe

“Classical” stochastic approximation

- **General problem of finding zeros of $h : \mathbb{R}^d \to \mathbb{R}^d$**
  - From random observations of values of h at certain points
  - Main example: minimization of $f : \mathbb{R}^d \to \mathbb{R}$, with $h = f'$

- **Classical algorithm (Robbins and Monro, 1951b)**

$$\theta_n = \theta_{n-1} - \gamma_n \big[h(\theta_{n-1}) + \varepsilon_n \big]$$

"Classical" stochastic approximation

- **General problem of finding zeros of $h : \mathbb{R}^d \to \mathbb{R}^d$**
  - From random observations of values of h at certain points
  - Main example: minimization of $f : \mathbb{R}^d \to \mathbb{R}$, with $h = f'$

- **Classical algorithm (Robbins and Monro, 1951b)**

$$\theta_n = \theta_{n-1} - \gamma_n \big[h(\theta_{n-1}) + \varepsilon_n \big]$$

- **Goals** (see, e.g., Duflo, 1996)
  - Beyond reducing noise by averaging observations
  - General sufficient conditions for convergence
  - Convergence in quadratic mean vs. convergence almost surely
  - Rates of convergences and choice of step-sizes
  - Asymptotics - no convexity

“Classical” stochastic approximation

- **Intuition from recursive mean estimation**
  - Starting from $\theta_0 = 0$, getting data $x_n \in \mathbb{R}^d$

$$\theta_n = \theta_{n-1} - \gamma_n(\theta_{n-1} - x_n)$$

  - If $\gamma_n = 1/n$, then $\theta_n = \frac{1}{n}\sum_{k=1}^n x_k$
  - If $\gamma_n = 2/(n+1)$ then $\theta_n = \frac{2}{n(n+1)}\sum_{k=1}^n kx_k$

"Classical" stochastic approximation

- **Intuition from recursive mean estimation**
  - Starting from $\theta_0 = 0$, getting data $x_n \in \mathbb{R}^d$

$$\theta_n = \theta_{n-1} - \gamma_n(\theta_{n-1} - x_n)$$

  - If $\gamma_n = 1/n$, then $\theta_n = \frac{1}{n}\sum_{k=1}^n x_k$
  - If $\gamma_n = 2/(n+1)$ then $\theta_n = \frac{2}{n(n+1)}\sum_{k=1}^n k x_k$

- In general: $\mathbb{E}x_n = x$ and thus $\theta_n - x = (1-\gamma_n)(\theta_{n-1}-x) + \gamma_n(x_n - x)$

$$\theta_n - x = \prod_{k=1}^n (1-\gamma_k)(\theta_0 - x) + \sum_{i=1}^n \prod_{k=i+1}^n (1-\gamma_k)\gamma_i(x_i - x)$$

“Classical” stochastic approximation

- Expanding the recursion with i.i.d. x_n's and $\sigma^2 = \mathbb{E}\|x_n - x\|^2$:

$$
\begin{aligned}
\theta_n - x &= \prod_{k=1}^{n}(1-\gamma_k)(\theta_0 - x) + \sum_{i=1}^{n}\gamma_i \prod_{k=i+1}^{n}(1-\gamma_k)(x_i - x) \\
\mathbb{E}\|\theta_n - x\|^2 &= \prod_{k=1}^{n}(1-\gamma_k)^2\|\theta_0 - x\|^2 + \sum_{i=1}^{n}\gamma_i^2 \prod_{k=i+1}^{n}(1-\gamma_k)^2\sigma^2
\end{aligned}
$$

"Classical" stochastic approximation

- Expanding the recursion with i.i.d. x_n's and $\sigma^2 = \mathbb{E}\|x_n - x\|^2$:

$$\theta_n - x = \prod_{k=1}^{n}(1-\gamma_k)(\theta_0 - x) + \sum_{i=1}^{n}\gamma_i \prod_{k=i+1}^{n}(1-\gamma_k)(x_i - x)$$

$$\mathbb{E}\|\theta_n - x\|^2 = \prod_{k=1}^{n}(1-\gamma_k)^2\|\theta_0 - x\|^2 + \sum_{i=1}^{n}\gamma_i^2 \prod_{k=i+1}^{n}(1-\gamma_k)^2\sigma^2$$

- Requires study of $\prod_{k=1}^{n}(1-\gamma_k)$ and $\sum_{i=1}^{n}\gamma_i^2\prod_{k=i+1}^{n}(1-\gamma_k)^2$
  - If $\gamma_n = o(1)$, $\log\prod_{k=1}^{n}(1-\gamma_k) \sim -\sum_{k=1}^{n}\gamma_k$ should go to $-\infty$
    Forgetting initial conditions (even arbitrarily far)
  - $\sum_{i=1}^{n}\gamma_i^2\prod_{k=i+1}^{n}(1-\gamma_k)^2 \sim \sum_{i=1}^{n}\gamma_i^2\prod_{k=i+1}^{n}(1-2\gamma_k)$
    Robustness to noise

Forgetting of initial conditions

$$\log \prod_{k=1}^{n}(1-\gamma_k) \;\sim\; -\sum_{k=1}^{n}\gamma_k$$

- Examples: $\boxed{\gamma_n = C/n^\alpha}$
  - $\alpha = 1$, $\sum_{i=1}^n \frac{1}{i} = \log(n) + \text{ cst } + O(1/n)$
  - $\alpha > 1$, $\sum_{i=1}^n \frac{1}{i^\alpha} = \text{ cst } + O(1/n^{\alpha-1})$
  - $\alpha \in (0,1)$, $\sum_{i=1}^n \frac{1}{i^\alpha} = \text{ cst } \times n^{1-\alpha} + O(1)$
  - Proof using relationship with integrals

- Consequences
  - if $\alpha > 1$, no convergence
  - If $\alpha \in (0,1)$, exponential convergence
  - if $\alpha = 1$, convergence of squared norm in $1/n^{2C}$

Decomposition of the noise term

- Assume (γ_n) is decreasing and less than 1; then for any $m \in \{1,\dots,n\}$, we may split the following sum as follows:

$$\begin{aligned}
\sum_{k=1}^{n}\prod_{i=k+1}^{n}(1-\gamma_i)\gamma_k^2 &= \sum_{k=1}^{m}\prod_{i=k+1}^{n}(1-\gamma_i)\gamma_k^2 + \sum_{k=m+1}^{n}\prod_{i=k+1}^{n}(1-\gamma_i)\gamma_k^2 \\
&\leqslant \prod_{i=m+1}^{n}(1-\gamma_i)\sum_{k=1}^{m}\gamma_k^2 + \gamma_m \sum_{k=m+1}^{n}\prod_{i=k+1}^{n}(1-\gamma_i)\gamma_k \\
&\leqslant \exp\left(-\sum_{i=m+1}^{n}\gamma_i\right)\sum_{k=1}^{m}\gamma_k^2 + \gamma_m \sum_{k=m+1}^{n}\left[\prod_{i=k+1}^{n}(1-\gamma_i) - \prod_{i=k}^{n}(1-\gamma_i)\right] \\
&\leqslant \exp\left(-\sum_{i=m+1}^{n}\gamma_i\right)\sum_{k=1}^{m}\gamma_k^2 + \gamma_m\left[1 - \prod_{i=m+1}^{n}(1-\gamma_i)\right] \\
&\leqslant \exp\left(-\sum_{i=m+1}^{n}\gamma_i\right)\sum_{k=1}^{n}\gamma_k^2 + \gamma_m
\end{aligned}$$

Decomposition of the noise term

$$\sum_{k=1}^{n} \prod_{i=k+1}^{n} (1-\gamma_i)\gamma_k^2 \leqslant \exp\left(-\sum_{i=m+1}^{n} \gamma_i\right) \sum_{k=1}^{n} \gamma_k^2 + \gamma_m$$

- Require γ_n to tend to zero (vanishing decaying step-size)
  - May not need $\sum_n \gamma_n^2 < \infty$ for convergence in quadratic mean

- Examples: $\boxed{\gamma_n = C/n^\alpha}$ and mean estimation, with $m = n/2$
  - No need to consider $\alpha > 1$
  - $\alpha \in (0,1)$, $\exp(-C' n^{1-\alpha}) n^{\max\{1-2\alpha,0\}} + O(Cn^{-\alpha})$
  - $\alpha = 1$, convergence of noise term in $O(1/n)$ but forgetting of initial condition in $O(1/n^{2C})$
  - Consequences: need $\alpha \in (0,1]$ and $C \geqslant 1/2$ for $\alpha = 1$

Robbins-Monro algorithm

- **General problem of finding zeros of** $h : \mathbb{R}^d \to \mathbb{R}^d$
  - From random observations of values of h at certain points
  - Main example: minimization of $f : \mathbb{R}^d \to \mathbb{R}$, with $h = f'$

- **Classical algorithm (Robbins and Monro, 1951b)**

$$\theta_n = \theta_{n-1} - \gamma_n \big[h(\theta_{n-1}) + \varepsilon_n \big]$$

- **Goals** (see, e.g., Duflo, 1996)
  - General sufficient conditions for convergence
  - Convergence in quadratic mean vs. convergence almost surely
  - Rates of convergences and choice of step-sizes
  - Asymptotics - no convexity

Different types of convergences

- **Goal**: show that $\theta_n \to \theta_*$ or $d(\theta_n, \Theta_*) \to 0$ or $f(\theta_n) \to f(\theta_*)$
  - Random quantity $\delta_n \in \mathbb{R}$ tending to zero
- **Convergence almost-surely**: $\mathbb{P}(\delta_n \to 0) = 1$
- **Convergence in probability**: $\forall \varepsilon > 0, \mathbb{P}(|\delta_n| \geqslant \varepsilon) \to 0$
- **Convergence in mean** $r \geqslant 1$: $\mathbb{E}|\delta_n|^r \to 0$

Different types of convergences

- **Goal**: show that $\theta_n \to \theta_*$ or $d(\theta_n, \Theta_*) \to 0$ or $f(\theta_n) \to f(\theta_*)$
  - Random quantity $\delta_n \in \mathbb{R}$ tending to zero

- **Convergence almost-surely**: $\mathbb{P}(\delta_n \to 0) = 1$

- **Convergence in probability**: $\forall \varepsilon > 0, \mathbb{P}(|\delta_n| \geqslant \varepsilon) \to 0$

- **Convergence in mean** $r \geqslant 1$: $\mathbb{E}|\delta_n|^r \to 0$

- **Relationship between convergences**
  - Almost surely $\Rightarrow$ in probability
  - In mean $\Rightarrow$ in probability (Markov's inequality)
  - In probability (sufficiently fast) $\Rightarrow$ almost surely (Borel-Cantelli)
  - Almost surely $+$ domination $\Rightarrow$ in mean

Robbins-Monro algorithm
Need for Lyapunov functions (even with no noise)

$$\theta_n = \theta_{n-1} - \gamma_n \big[h(\theta_{n-1}) + \varepsilon_n \big]$$

- The Robbins-Monro algorithm cannot converge all the time...

- **Lyapunov function** $V : \mathbb{R}^d \to \mathbb{R}$ with following properties
  - Non-negative values: $V \geqslant 0$
  - Continuously-differentiable with L-Lipschitz-continuous gradients
  - Control of h: $\forall \theta$, $\|h(\theta)\|^2 \leqslant C(1 + V(\theta))$
  - Gradient condition: $\forall \theta$, $\boxed{h(\theta)^\top V'(\theta) \geqslant \alpha \|V'(\theta)\|^2}$

Robbins-Monro algorithm
Need for Lyapunov functions (even with no noise)

$$\theta_n = \theta_{n-1} - \gamma_n \big[h(\theta_{n-1}) + \varepsilon_n \big]$$

- The Robbins-Monro algorithm cannot converge all the time...
- **Lyapunov function** $V : \mathbb{R}^d \to \mathbb{R}$ with following properties
  - Non-negative values: $V \geqslant 0$
  - Continuously-differentiable with L-Lipschitz-continuous gradients
  - Control of h: $\forall \theta$, $\|h(\theta)\|^2 \leqslant C(1 + V(\theta))$
  - Gradient condition: $\forall \theta$, $\boxed{h(\theta)^\top V'(\theta) \geqslant \alpha' \|V'(\theta)\|^2}$
- If $h = f'$, then $V(\theta) = f(\theta) - \inf f$ is the default (but not only) choice for Lyapunov function: applies also to non-convex functions
  - Will require often some additional condition $\|V'(\theta)\|^2 \geqslant 2\mu V(\theta)$

Robbins-Monro algorithm
Martingale noise

$$\theta_n = \theta_{n-1} - \gamma_n \big[h(\theta_{n-1}) + \varepsilon_n \big]$$

- **Assumptions about the noise** ε_n
  - Typical assumption: ε_n i.i.d. $\Rightarrow$ not needed
  - "information up to time n": sequence of increasing σ-fields $\mathcal{F}_n$
  - Example from machine learning: $\mathcal{F}_n = \sigma(x_1, y_1, \dots, x_n, y_n)$
  - Assume $\boxed{\mathbb{E}(\varepsilon_n | \mathcal{F}_{n-1}) = 0}$ and $\boxed{\mathbb{E}\big[\|\varepsilon_n\|^2 | \mathcal{F}_{n-1}\big] \leqslant \sigma^2}$ almost surely

- Warning: SGD for machine learning does not correspond to ε_n i.i.d.

- **Key property**: θ_n is $\mathcal{F}_n$-measurable

Robbins-Monro algorithm
Convergence of the Lyapunov function

- Using regularity (and other properties) of V:

$$
\begin{aligned}
V(\theta_n) &\leqslant V(\theta_{n-1}) + V'(\theta_{n-1})^\top(\theta_n - \theta_{n-1}) + \frac{L}{2}\|\theta_n - \theta_{n-1}\|^2 \\
&= V(\theta_{n-1}) - \gamma_n V'(\theta_{n-1})^\top\big(h(\theta_{n-1}) + \varepsilon_n\big) + \frac{L\gamma_n^2}{2}\|h(\theta_{n-1}) + \varepsilon_n\|^2 \\
\mathbb{E}\big[V(\theta_n)|\mathcal{F}_{n-1}\big] &\leqslant V(\theta_{n-1}) - \gamma_n V'(\theta_{n-1})^\top h(\theta_{n-1}) + \frac{L\gamma_n^2}{2}\|h(\theta_{n-1})\|^2 + \frac{L\gamma_n^2}{2}\sigma^2 \\
&\leqslant V(\theta_{n-1}) - \alpha'\gamma_n\|V'(\theta_{n-1})\|^2 + \frac{LC\gamma_n^2}{2}\big[1 + V(\theta_{n-1})\big] + \frac{L\gamma_n^2}{2}\sigma^2 \\
&\leqslant V(\theta_{n-1})\big[1 + \frac{LC\gamma_n^2}{2}\big] - \alpha'\gamma_n\|V'(\theta_{n-1})\|^2 + \frac{L\gamma_n^2}{2}(C + \sigma^2)
\end{aligned}
$$

Robbins-Monro algorithm
Convergence of the expected Lyapunov function
with "curvature"

- If $\|V'(\theta)\|^2 \geqslant 2\mu V(\theta)$ and $\gamma_n \leqslant \frac{2\alpha'\mu}{LC}$:

$$\begin{aligned} \mathbb{E}\big[V(\theta_n)|\mathcal{F}_{n-1}\big] &\leqslant V(\theta_{n-1})\big[1-\alpha'\mu\gamma_n\big] + M\gamma_n^2 \\ \mathbb{E}V(\theta_n) &\leqslant \mathbb{E}V(\theta_{n-1})\big[1-\alpha'\mu\gamma_n\big] + M\gamma_n^2 \end{aligned}$$

- Need to study non-negative sequence $\delta_n \leqslant \delta_{n-1}\big[1-\alpha'\mu\gamma_n\big] + M\gamma_n^2$ with $\delta_n = \mathbb{E}V(\theta_n)$

- Sufficient conditions for convergence of the expected Lyapunov function (with curvature)
  - $\sum_n \gamma_n = +\infty$ and $\gamma_n \to 0$
  - Special case of $\gamma_n = C/n^\alpha$

Robbins-Monro algorithm
Convergence of the expected Lyapunov function
with "curvature" - $\gamma_n = C/n^\alpha$

- Need to study non-negative sequence $\delta_n \leqslant \delta_{n-1}\big[1 - \alpha'\mu\gamma_n\big] + M\gamma_n^2$ with $\delta_n = \mathbb{E}V(\theta_n)$ (NB: forgetting constraint on γ_n - see next class)

$$\delta_n \leqslant \prod_{k=1}^{n}(1-\alpha'\mu\gamma_k)\delta_0 + M\sum_{i=1}^{n}\gamma_i^2 \prod_{k=i+1}^{n}(1-\alpha'\mu\gamma_k)$$

- If $\alpha > 1$: no forgetting of initial conditions

- If $\alpha \in (0,1)$: $\delta_0 \exp(-$ cst $\alpha'\mu C \times n^{1-\alpha}) + \gamma_n M$

- If $\alpha = 1$ and $\gamma_n = C/n$: $\delta_0 n^{-\mu C} + \gamma_n M$

Robbins-Monro algorithm
Almost-sure convergence

- Using regularity of V:

$$
\begin{aligned}
V(\theta_n) &\leqslant V(\theta_{n-1}) + V'(\theta_{n-1})^\top(\theta_n - \theta_{n-1}) + \frac{L}{2}\|\theta_n - \theta_{n-1}\|^2 \\
&= V(\theta_{n-1}) - \gamma_n V'(\theta_{n-1})^\top(h(\theta_{n-1}) + \varepsilon_n) + \frac{L\gamma_n^2}{2}\|h(\theta_{n-1}) + \varepsilon_n\|^2 \\
\mathbb{E}\big[V(\theta_n)|\mathcal{F}_{n-1}\big] &\leqslant V(\theta_{n-1}) - \gamma_n V'(\theta_{n-1})^\top h(\theta_{n-1}) + \frac{L\gamma_n^2}{2}\|h(\theta_{n-1})\|^2 + \frac{L\gamma_n^2}{2}\sigma^2 \\
&\leqslant V(\theta_{n-1}) - \alpha'\gamma_n\|V'(\theta_{n-1})\|^2 + \frac{LC\gamma_n^2}{2}\big[1 + V(\theta_{n-1})\big] + \frac{L\gamma_n^2}{2}\sigma^2 \\
&= V(\theta_{n-1})\big[1 + \frac{LC\gamma_n^2}{2}\big] - \alpha'\gamma_n\|V'(\theta_{n-1})\|^2 + \frac{L\gamma_n^2}{2}(C + \sigma^2)
\end{aligned}
$$

Robbins and Siegmund (1985)

- **Assumptions**
  - Measurability: Let V_n, β_n, χ_n, η_n four $\mathcal{F}_n$-adapted real sequences
  - Non-negativity: V_n, β_n, χ_n, η_n non-negative
  - Summability: $\sum_n \beta_n < \infty$ and $\sum_n \chi_n < \infty$
  - Inequality: $\mathbb{E}\big[V_n | \mathcal{F}_{n-1}\big] \leqslant V_{n-1}(1+\beta_{n-1}) + \chi_{n-1} - \eta_{n-1}$

- **Theorem**: (V_n) converges almost surely to a random variable V_∞ and $\sum_n \eta_n$ is finite almost surely

- *Proof*

- Consequence for stochastic approximation (if $\|V'(\theta)\|^2 \geqslant 2\mu V(\theta)$): $V(\theta_n)$ and $\|V'(\theta_n)\|^2$ converges almost surely to zero

Robbins and Siegmund (1985) - Proof sketch

- Inequality: $\mathbb{E}\big[V_n | \mathcal{F}_{n-1}\big] \leqslant V_{n-1}(1+\beta_{n-1}) + \chi_{n-1} - \eta_{n-1}$

- Define $\alpha_n = \prod_{k=1}^n (1+\beta_k)$ a converging sequence, $V_n' = \alpha_{n-1} V_n$, $\chi_n' = \alpha_{n-1}\chi_n$ and $\eta_n' = \alpha_{n-1}\eta_n$ so that:

$$\mathbb{E}\big[V_n' | \mathcal{F}_{n-1}\big] \leqslant V_{n-1} + \chi_{n-1}' - \eta_{n-1}'$$

- Define the super-martingale $Y_n = V_n' - \sum_{k=1}^{n-1} (\chi_k' - \eta_k')$ so that $\mathbb{E}\big[Y_n | \mathcal{F}_{n-1}\big] \leqslant Y_{n-1}$

- Probabilistic proof using Doob convergence theorem (Duflo, 1996)

Robbins-Monro analysis - non random errors

- **Random unbiased errors**: no need for vanishing magnitudes

- **Non-random errors**: need for vanishing magnitudes
  - See Duflo (1996, Theorem 2.III.4)
  - See also Schmidt et al. (2011)

Robbins-Monro analysis - asymptotic normality (Fabian, 1968)

- Traditional step-size $\gamma = C/n$ (and proof sketch for differential A of h at unique θ_* symmetric)

$$\begin{aligned}
\theta_n &= \theta_{n-1} - \gamma_n h(\theta_{n-1}) - \gamma_n \varepsilon_n \\
&\approx \theta_{n-1} - \gamma_n \big[h'(\theta_*)(\theta_{n-1} - \theta_*)\big] - \gamma_n \varepsilon_n + \gamma_n O(\|\theta_n - \theta_*\|^2) \\
&\approx \theta_{n-1} - \gamma_n A(\theta_{n-1} - \theta_*) - \gamma_n \varepsilon_n \\
\theta_n - \theta_* &\approx (I - \gamma_n A) \cdots (I - \gamma_1 A)(\theta_0 - \theta_*) - \sum_{k=1}^n (I - \gamma_n A) \cdots (I - \gamma_{k+1} A) \gamma_k \varepsilon_k \\
\theta_n - \theta_* &\approx \exp\big[-(\gamma_n + \cdots + \gamma_1)A\big](\theta_0 - \theta_*) - \sum_{k=1}^n \exp\big[-(\gamma_n + \cdots + \gamma_{k+1})A\big]\gamma_k \varepsilon_k \\
&\approx \exp\big[-CA\log n\big](\theta_0 - \theta_*) - \sum_{k=1}^n \exp\big[-C(\log n - \log k)A\big]\frac{C}{k}\varepsilon_k
\end{aligned}$$

- Asymptotic normality by averaging random variables

Robbins-Monro analysis - asymptotic normality (Fabian, 1968)

- Assuming A, $(\theta_0 - \theta_*)(\theta_0 - \theta_*)^\top$ and $\mathbb{E}(\varepsilon_k \varepsilon_k^\top) = \Sigma$ commute

$$\theta_n - \theta_* \approx \exp\big[-CA\log n\big](\theta_0 - \theta_*) - \sum_{k=1}^n \exp\big[-C(\log n - \log k)A\big]\frac{C}{k}\varepsilon_k$$

$$\begin{aligned}
\mathbb{E}(\theta_n - \theta_*)(\theta_n - \theta_*)^\top &\approx \exp\big[-2CA\log n\big](\theta_0 - \theta_*)(\theta_0 - \theta_*)^\top \\
&\quad + \sum_{k=1}^n \exp\big[-2C(\log n - \log k)A\big]\frac{C^2}{k^2}\mathbb{E}(\varepsilon_k \varepsilon_k^\top) \\
&\approx n^{-2CA}(\theta_0 - \theta_*)(\theta_0 - \theta_*)^\top + n^{-2CA}\sum_{k=1}^n C^2 k^{2CA-2}\Sigma \\
&\approx n^{-2CA}(\theta_0 - \theta_*)(\theta_0 - \theta_*)^\top + n^{-2CA} C^2 \frac{n^{2CA-1}}{2CA-1}\Sigma
\end{aligned}$$

Robbins-Monro analysis - asymptotic normality (Fabian, 1968)

$$\mathbb{E}(\theta_n - \theta_*)(\theta_n - \theta_*)^\top \approx n^{-2CA}(\theta_0 - \theta_*)(\theta_0 - \theta_*)^\top + \frac{1}{n}C^2\frac{1}{2CA - 1}\Sigma$$

- Step-size $\gamma = C/n$ (note that this only a sketch of proof)
  - Need $2C\lambda_{\min}(A) \geqslant 1$ for convergence, which implies that the first term depending on initial condition $\theta_* - \theta_0$ is negligible
  - C too small $\Rightarrow$ no convergence - C too large $\Rightarrow$ large variance

- Dependence on the conditioning of the problem
  - If $\lambda_{\min}(A)$ is small, then C is large
  - "Choosing" A proportional to identity for optimal behavior (by premultiplying A by a conditioning matrix that make A close to a constant times identity

Polyak-Ruppert averaging

- **Problems with Robbins-Monro algorithm**
  - Choice of step-sizes in Robbins-Monro algorithm
  - Dependence on the unknown conditioning of the problem

- **Simple but impactful idea** (Polyak and Juditsky, 1992; Ruppert, 1988)
  - Consider the averaged iterate $\boxed{\bar{\theta}_n = \frac{1}{n}\sum_{k=1}^{n}\theta_k}$
  - NB: "Offline" averaging
  - Can be computed recursively as $\bar{\theta}_n = (1-1/n)\bar{\theta}_{n-1} + \frac{1}{n}\theta_n$
  - In practice, may start the averaging "after a while"

- **Analysis**
  - Unique optimum θ_*. See details by Polyak and Juditsky (1992)

Cesaro means

- Assume $\theta_n \to \theta_*$, with convergence rate $\|\theta_n - \theta_*\| \leqslant \alpha_n$
- Cesaro's theorem: $\bar{\theta}_n = \frac{1}{n}\sum_{k=1}^{n} \theta_n$ converges to θ_*
- What about convergence rate $\|\bar{\theta}_n - \theta_*\|$?

Cesaro means

- Assume $\theta_n \to \theta_*$, with convergence rate $\|\theta_n - \theta_*\| \leqslant \alpha_n$
- Cesaro's theorem: $\bar{\theta}_n = \frac{1}{n}\sum_{k=1}^n \theta_n$ converges to θ_*
- What about convergence rate $\|\bar{\theta}_n - \theta_*\|$?

$$\|\bar{\theta}_n - \theta_*\| \;\leqslant\; \frac{1}{n}\sum_{k=1}^n \|\theta_k - \theta_*\| \leqslant \frac{1}{n}\sum_{k=1}^n \alpha_k$$

  - Will depend on rate α_n
  - If $\sum_n \alpha_n < \infty$, the rate becomes $1/n$ independently of α_n

Polyak-Ruppert averaging - Proof sketch - I

- Recursion: $\theta_n = \theta_{n-1} - \gamma_n(h(\theta_{n-1}) + \varepsilon_n)$ with $\gamma_n = C/n^\alpha$
  - From before, we know that $\|\theta_n - \theta_*\|^2 = O(n^{-\alpha})$

$$
\begin{aligned}
h(\theta_{n-1}) &= \frac{1}{\gamma_n}\big[\theta_{n-1} - \theta_n\big] - \varepsilon_n \\
A(\theta_{n-1} - \theta_*) + O(\|\theta_{n-1} - \theta_*\|^2) &= \frac{1}{\gamma_n}\big[\theta_{n-1} - \theta_n\big] - \varepsilon_n \text{ with } A = h'(\theta_*) \\
A(\theta_{n-1} - \theta_*) &= \frac{1}{\gamma_n}\big[\theta_{n-1} - \theta_n\big] - \varepsilon_n + O(n^{-\alpha}) \\
\frac{1}{n}\sum_{k=1}^{n} A(\theta_{k-1} - \theta_*) &= \frac{1}{n}\sum_{k=1}^{n}\frac{1}{\gamma_k}\big[\theta_{k-1} - \theta_k\big] - \frac{1}{n}\sum_{k=1}^{n}\varepsilon_k + O(n^{-\alpha}) \\
\frac{1}{n}\sum_{k=1}^{n} A(\theta_{k-1} - \theta_*) &= \frac{1}{n}\sum_{k=1}^{n}\frac{1}{\gamma_k}\big[\theta_{k-1} - \theta_k\big] + \text{Normal}(0, \Sigma/n) + O(n^{-\alpha})
\end{aligned}
$$

Polyak-Ruppert averaging - Proof sketch - II

- **Goal**: Bounding $\frac{1}{n}\sum_{k=1}^{n}\frac{1}{\gamma_k}\big[\theta_{k-1}-\theta_k\big]$ given $\|\theta_n-\theta_*\|^2 = O(n^{-\alpha})$

- Abel's summation formula: We have, summing by parts,

$$\frac{1}{n}\sum_{k=1}^{n}\frac{1}{\gamma_k}(\theta_{k-1}-\theta_k) = \frac{1}{n}\sum_{k=1}^{n-1}(\theta_k-\theta_*)(\gamma_{k+1}^{-1}-\gamma_k^{-1}) - \frac{1}{n}(\theta_n-\theta_*)\gamma_n^{-1} + \frac{1}{n}(\theta_0-\theta_*)\gamma_1^{-1}$$

leading to

$$\bigg\|\frac{1}{n}\sum_{k=1}^{n}\frac{1}{\gamma_k}(\theta_{k-1}-\theta_k)\bigg\| \leqslant \frac{1}{n}\sum_{k=1}^{n-1}\|\theta_k-\theta_*\|\cdot|\gamma_{k+1}^{-1}-\gamma_k^{-1}| + \frac{1}{n}\|\theta_n-\theta_*\|\gamma_n^{-1} + \frac{1}{n}\|\theta_0-\theta_*\|\gamma_1^{-1}$$

which is negligible

Polyak-Ruppert averaging - Proof sketch - III

- Recursion: $\theta_n = \theta_{n-1} - \gamma_n(h(\theta_{n-1}) + \varepsilon_n)$ with $\gamma_n = C/n^\alpha$
  - From before, we know that $\|\theta_n - \theta_*\|^2 = O(n^{-\alpha})$

$$\frac{1}{n}\sum_{k=1}^{n} A(\theta_{k-1} - \theta_*) \quad = \quad \text{Normal}(0, \Sigma/n) + O(n^{-\alpha}) + O(n^{2\alpha-1})$$

- **Consequence**: $\bar{\theta}_n - \theta_*$ is asymptotically normal with mean zero and covariance $\frac{1}{n}A^{-1}\Sigma A^{-1}$
  - Achieves the Cramer-Rao lower bound (see next lecture)
  - Independent of step-size (see next lecture)
  - Where are the initial conditions? (see next lecture)

Beyond the classical analysis

- **Lack of strong-convexity**
  - Step-size $\gamma_n = 1/n$ not robust to ill-conditioning
- **Robustness of step-sizes**
- **Explicit forgetting of initial conditions**