MATHEMATICS OF DEEP LEARNING

JOAN BRUNA , CIMS + CDS, NYU, SPRING'18

Lecture 8: Discrete vs Continuous time in Optimization: the Stochastic case and beyond.

OBJECTIVES LECTURE 8

- Bregman Divergence (finalize lecture 7)

- Markov Chains and SGD

- Stochastic Modified Equations

- Stochastic Gradient Langevin Dynamics

ODE AND NESTEROV

- Substituting on the finite-difference equation, we obtain

$$\dot{X}(t) + \frac{1}{2}\ddot{X}(t)\sqrt{s} + o(\sqrt{s}) = \left(1 - \frac{3\sqrt{s}}{t}\right)\left(\dot{X}(t) - \frac{1}{2}\ddot{X}(t)\sqrt{s} + o(\sqrt{s})\right) - \sqrt{s}\nabla g(X(t)) + o(\sqrt{s}) \ .$$

- The coefficients in $\sqrt{s}$ thus yield

$$\ddot{X} + \frac{3}{t}\dot{X} + \nabla g(X) = 0 \ .$$

- Initial conditions:
  - $X(0) = \theta_0 \ ,$
  - $\dot{X}(0) = \lim_{\sqrt{s}\to 0} \frac{\theta_2 - \theta_1}{\sqrt{s}} = -\lim_{\sqrt{s}\to 0} \sqrt{s}\nabla g(\eta_1) = 0 \ .$

ODE AND NESTEROV

- Denote $\mathcal{F}_L$: Class of L-smooth convex functions.
- **Theorem [Su et al.14]:** For any $g \in \mathcal{F}_L$ and any $\theta_0 \in \mathbb{R}^n$, the previous ODE with init condition $X(0) = \theta_0$, $\dot{X}(0) = 0$ has a unique global solution $X \in \mathcal{C}^2((0,\infty), \mathbb{R}^n) \cap \mathcal{C}^1([0,\infty), \mathbb{R}^n)$.
  - The ODE is well-posed despite being singular at $t = 0$.

ODE AND NESTEROV

- Denote $\mathcal{F}_L$: Class of L-smooth convex functions.
- **Theorem [Su et al.14]:** For any $g \in \mathcal{F}_L$ and any $\theta_0 \in \mathbb{R}^n$, the previous ODE with init condition $X(0) = \theta_0$, $\dot{X}(0) = 0$ has a unique global solution $X \in \mathcal{C}^2((0,\infty), \mathbb{R}^n) \cap \mathcal{C}^1([0,\infty), \mathbb{R}^n)$.
  - The ODE is well-posed despite being singular at $t = 0$.
- **Theorem [Su et al.14]:** For any $g \in \mathcal{F}_L$, as $s \to 0$, Nesterov scheme converges to the previous ODE: for all $T > 0$,

$$\lim_{s \to 0} \max_{k \leq T/\sqrt{s}} \|\theta_k - X(k\sqrt{s})\| = 0 \ .$$

  - Nesterov is a proper discretization.
  - Despite being a first-order method, ODE is 2nd order.

CONSEQUENCES: ANALOGOUS CONVERGENCE RATE

- Recall that discrete Nesterov scheme with step s satisfies

$$g(\theta_k) - g(\theta_*) \leq \frac{2\|\theta_0 - \theta_*\|^2}{s(k+1)^2}$$

- The continuous ODE satisfies similar quadratic rate:

Theorem [Su et al'14]: For any $g \in \mathcal{F}_L$, let $X(t)$ be the unique solution of the ODE with initial conditions $X(0) = \theta_0$ and $\dot{X}(0) = 0$. Then

$$g(X(t)) - g(\theta_*) \leq \frac{2\|\theta_0 - \theta_*\|^2}{t^2} \ .$$

- Consistent with the sampling rate $t = k\sqrt{s}$.

NESTEROV VS GRADIENT DESCENT

- The equivalence $t \approx k\sqrt{s}$ suggests that running Nesterov's scheme for different learning rates $s, s' (\leq L^{-1})$ will produce iterates $(\theta_k)_k, (\theta'_{k'})_{k'}$ such that

$$\theta_k \approx \theta'_{k'} \text{ if } k\sqrt{s} \approx k'\sqrt{s'} \ .$$

NESTEROV VS GRADIENT DESCENT

- The equivalence $t \approx k\sqrt{s}$ suggests that running Nesterov's scheme for different learning rates $s, s' (\leq L^{-1})$ will produce iterates $(\theta_k)_k, (\theta'_{k'})_{k'}$ such that

$$\theta_k \approx \theta'_{k'} \text{ if } k\sqrt{s} \approx k'\sqrt{s'} \ .$$

- The integral curve that solves the ODE "contains" discrete iterates for different step sizes.
- Each Nesterov iteration travels $\sqrt{s}$ units of time.

NESTEROV VS GRADIENT DESCENT

- The equivalence $t \approx k\sqrt{s}$ suggests that running Nesterov's scheme for different learning rates $s, s' (\leq L^{-1})$ will produce iterates $(\theta_k)_k, (\theta'_{k'})_{k'}$ such that

$$\theta_k \approx \theta'_{k'} \text{ if } k\sqrt{s} \approx k'\sqrt{s'} \ .$$

- The integral curve that solves the ODE "contains" discrete iterates for different step sizes.
- Each Nesterov iteration travels $\sqrt{s}$ units of time.
- In contrast, each Gradient descent step moves s units of time along integral cuve of gradient flow $\dot{X} + \nabla g(X) = 0$.

CONSTANT 3?

- What is the role of the constant 3 in the ODE

$$\ddot{X} + \frac{3}{t}\dot{X} + \nabla g(X) = 0 \ .$$

CONSTANT 3?

- What is the role of the constant 3 in the ODE

$$\ddot{X} + \frac{3}{t}\dot{X} + \nabla g(X) = 0 \ .$$

- **Theorem [Su et al'14]:** If $r \geq 3$, the solution of $\ddot{X} + \frac{r}{t}\dot{X} + \nabla g(X) = 0$ satisfies

$$g(X(t)) - g(\theta^*) \leq \frac{(r-1)^2 \|\theta_0 - \theta_*\|^2}{2t^2} \ .$$

- If $r < 3$, inverse quadratic rate is lost: $r \geq 3$ is necessary and sufficient for quadratic rate. Phase transition.

FROM CONTINUOUS TO DISCRETE TIME

- The previous work derived a continuous-time ODE that is the limit of a given discrete optimization scheme.

FROM CONTINUOUS TO DISCRETE TIME

- The previous work derived a continuous-time ODE that is the limit of a given discrete optimization scheme.
- Can we take the opposite route?

FROM CONTINUOUS TO DISCRETE TIME

- The previous work derived a continuous-time ODE that is the limit of a given discrete optimization scheme.
- Can we take the opposite route?

- Given convex g, a general procedure to yield acceleration:

1. Construct and ODE such that its solutions $X(t)$ satisfy
$$g(X(t)) - g(\theta_*) \lesssim t^{-p}$$

2. Discretize the ODE in such a way to preserve convergence.

THE BREGMAN LAGRANGIAN [WIBISONO, WILSON, JORDAN]

- Assume the minimisation $\min_{\theta \in \Theta} g(\theta)$ admits a unique minimiser θ_* .

THE BREGMAN LAGRANGIAN [WIBISONO, WILSON, JORDAN]

- Assume the minimisation $\min_{\theta \in \Theta} g(\theta)$ admits a unique minimiser θ_* .
- Consider an auxiliary convex function h, to define a notion of distance in Θ via the *Bregman divergence*:

$$D_h(\theta, \eta) = h(\theta) - h(\eta) - \langle \nabla h(\eta), \theta - \eta \rangle$$

THE BREGMAN LAGRANGIAN [WIBISONO, WILSON, JORDAN]

- Assume the minimisation $\min_{\theta \in \Theta} g(\theta)$ admits a unique minimiser θ_* .
- Consider an auxiliary convex function h, to define a notion of distance in Θ via the *Bregman divergence*:

$$D_h(\theta, \eta) = h(\theta) - h(\eta) - \langle \nabla h(\eta), \theta - \eta \rangle$$

- non-negative thanks to convexity of h .
- locally equivalent to the Hessian metric

$$D_h(\theta, \eta) = \frac{1}{2}(y - x)^\top \nabla^2 h(x)(y - x) + o(\|y - x\|)$$

THE BREGMAN LAGRANGIAN

- The *Bregman Lagrangian* is defined as

$$\mathcal{L}(X, V, t) := e^{\alpha(t)+\gamma(t)} \left(D_h(X + e^{-\alpha(t)} V, X) - e^{\beta(t)} g(X) \right) .$$

X: position, V: velocity, t: time.

THE BREGMAN LAGRANGIAN

- The *Bregman Lagrangian* is defined as

$$\mathcal{L}(X, V, t) := e^{\alpha(t)+\gamma(t)} \left(D_h(X + e^{-\alpha(t)} V, X) - e^{\beta(t)} g(X) \right) .$$

X: position, V: velocity, t: time.

- The functions $\alpha(t), \gamma(t), \beta(t)$ satisfy *ideal scaling* if

$$\dot{\beta}(t) \leq e^{\alpha(t)} \ , \ \dot{\gamma}(t) = e^{\alpha(t)}.$$

THE BREGMAN LAGRANGIAN

- The *Bregman Lagrangian* is defined as

$$\mathcal{L}(X, V, t) := e^{\alpha(t)+\gamma(t)} \left(D_h(X + e^{-\alpha(t)} V, X) - e^{\beta(t)} g(X) \right) .$$

X: position, V: velocity, t: time.

- The functions $\alpha(t), \gamma(t), \beta(t)$ satisfy *ideal scaling* if

$$\dot{\beta}(t) \leq e^{\alpha(t)} \ , \ \dot{\gamma}(t) = e^{\alpha(t)}.$$

- Given a path in space-time $X_t \in \Theta; t \geq 0$, the *action* is obtained by integrating the Lagrangian of the system:

$$\mathcal{J}(X) = \int_{t \geq 0} \mathcal{L}(X_t, \dot{X}_t, t) dt \, .$$

THE BREGMAN LAGRANGIAN

➤ The minimal action curves necessarily satisfy the *Euler-Lagrange equation*:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathcal{L}}{\partial V} = \frac{\partial \mathcal{L}}{\partial X}\,.$$

THE BREGMAN LAGRANGIAN

- The minimal action curves necessarily satisfy the *Euler-Lagrange equation*:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathcal{L}}{\partial V} = \frac{\partial \mathcal{L}}{\partial X}\,.$$

- In the case of the Bregman Lagrangian with ideal scaling, this equation becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla h(X_t + e^{-\alpha(t)}\dot{X}_t) = -e^{\alpha(t)+\beta(t)}\nabla g(X_t)\,.$$

THE BREGMAN LAGRANGIAN

- The minimal action curves necessarily satisfy the *Euler-Lagrange equation*:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathcal{L}}{\partial V} = \frac{\partial \mathcal{L}}{\partial X}\,.$$

- In the case of the Bregman Lagrangian with ideal scaling, this equation becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla h(X_t + e^{-\alpha(t)}\dot{X}_t) = -e^{\alpha(t)+\beta(t)}\nabla g(X_t)\,.$$

- Do the solutions minimize g ? How fast?

THE BREGMAN LAGRANGIAN

- **Theorem [Wibisono et al'17]:** Under ideal scaling, the solutions to the Euler-Lagrange equation satisfy

$$g(X_t) - g(\theta_*) \leq O(e^{-\beta(t)}) \,.$$

THE BREGMAN LAGRANGIAN

- **Theorem [Wibisono et al'17]:** Under ideal scaling, the solutions to the Euler-Lagrange equation satisfy

$$g(X_t) - g(\theta_*) \leq O(e^{-\beta(t)}) \,.$$

- Consider the Lyapunov function

$$\mathcal{E}_t = D_h(\theta_*, X_t + e^{-\alpha(t)}\dot{X}_t) + e^{\beta(t)}\left(g(X_t) - g(\theta_*)\right) \,.$$

- For a given $\alpha(t)$, the optimal convergence rate is achieved by $\dot{\beta}(t) = e^{\alpha(t)}$.

THE BREGMAN LAGRANGIAN

- **Theorem [Wibisono et al'17]:** Under ideal scaling, the solutions to the Euler-Lagrange equation satisfy

$$g(X_t) - g(\theta_*) \leq O(e^{-\beta(t)}) \,.$$

- Consider the Lyapunov function

$$\mathcal{E}_t = D_h(\theta_*, X_t + e^{-\alpha(t)}\dot{X}_t) + e^{\beta(t)}\left(g(X_t) - g(\theta_*)\right) \,.$$

- For a given $\alpha(t)$, the optimal convergence rate is achieved by $\dot{\beta}(t) = e^{\alpha(t)}$.

- We want fast convergence rate, but such that discretization preserves the rate!

THE BREGMAN LAGRANGIAN

➤ Consider the Bregman Lagrangians generated by parameters

$$\alpha(t) = \log p - \log t\,,\ \beta(t) = p\log t + \log C\,,\ \gamma(t) = p\log t\,.$$
$$(p > 0, C > 0)\ .$$

THE BREGMAN LAGRANGIAN

- Consider the Bregman Lagrangians generated by parameters

$$\alpha(t) = \log p - \log t\,,\ \beta(t) = p\log t + \log C\,,\ \gamma(t) = p\log t\,.$$
$$(p > 0, C > 0)\ .$$

- They satisfy the ideal scaling condition, and the resulting Euler-Lagrange equation is

$$\ddot{X}_t + \frac{p+1}{t}\dot{X}_t + Cp^2 t^{p-2}\left[\nabla^2 h(X_t + tp^{-1}\dot{X}_t)\right]^{-1}\nabla g(X_t) = 0\,.$$

  - From previous theorem, convergence rate $O(t^{-p})$.
  - $p = 2$ and $h(x) = \frac{1}{2}\|x\|^2$ is the ODE from Su et al.

DISCRETIZING EULER LAGRANGE EQUATIONS

- The decoupled first-order system of equations is

$$Z_t = X_t + \frac{t}{p}\dot{X}_t \,,\ \frac{d}{dt}\nabla h(Z_t) = -Cpt^{p-1}\nabla g(X_t) \ .$$

- Discrete Euler scheme becomes

$$\eta_k = \arg\min_z \left\{ Cpk^{p-1}\langle \nabla g(\theta_k), z\rangle + \delta^{-p} D_h(z, \eta_{k-1}) \right\}$$

$$\theta_{k+1} = \frac{p}{k}\eta_k + \frac{k-p}{k}\theta_k$$

DISCRETIZING EULER LAGRANGE EQUATIONS

- The decoupled first-order system of equations is

$$Z_t = X_t + \frac{t}{p}\dot{X}_t \,,\ \frac{d}{dt}\nabla h(Z_t) = -Cpt^{p-1}\nabla g(X_t) \ .$$

- Discrete Euler scheme becomes

$$\eta_k = \arg\min_z \left\{ Cpk^{p-1}\langle \nabla g(\theta_k), z\rangle + \delta^{-p} D_h(z, \eta_{k-1}) \right\}$$

$$\theta_{k+1} = \frac{p}{k}\eta_k + \frac{k-p}{k}\theta_k$$

- It turns out that the simple forward-backward Euler method to discretize this ODE is not stable, thus does not produce a discrete optimization scheme with matching convergence rate.

DISCRETIZING EULER-LAGRANGE EQUATIONS

- Nesterov's constructions on non-Euclidean domains (mirror descent) provide a general stable discretization scheme with matching rates.

$$\theta_{k+1} = \frac{p}{k+p}\eta_k + \frac{k}{k+p}\xi_k$$

$$\eta_k = \arg\min_z \left\{ Cpk^{p-1}\langle \nabla g(\xi_k), z\rangle + \delta^{-p} D_h(z, \eta_{k-1}) \right\}$$

with ξ_k satisfying

$$\langle \nabla g(\xi_k), \theta_k - \xi_k\rangle \geq M\delta^{p/(p-1)} \|\nabla g(\xi_k)\|^{p/(p-1)} .$$

- generalization of co-coercivity property.
- explicit construction for high-order accelerated gradients.

FROM CONTINUOUS TO DISCRETE TIME

- Other related works:
  - Lyapunov Analysis [Wilson, Recht, Jordan,'17]
  - Symplectic Optimization [Bettencourt, Wilson, Jordan]

- Extensions:
  - Composite optimization with proximal methods.
  - non-Euclidean settings: mirror descent, accelerated mirror descent.

LECTURE 8

- Bregman Divergence (finalize lecture 7)

- Markov Chains and SGD

- Stochastic Modified Equations

- Stochastic Gradient Langevin Dynamics

STOCHASTIC GRADIENT DESCENT

- Consider as before the minimization of a function g defined on $\mathbb{R}^d$.
- However, now we cannot compute $\nabla g(\theta)$ directly: only given access to unbiased estimates $\nabla g_n(\theta_n)$ at certain points θ_n.

STOCHASTIC GRADIENT DESCENT

- Consider as before the minimization of a function g defined on $\mathbb{R}^d$.
- However, now we cannot compute $\nabla g(\theta)$ directly: only given access to unbiased estimates $\nabla g_n(\theta_n)$ at certain points θ_n.
- In our setup, g_n is the loss for a single data-point:

$$g_n(\theta) = \ell(\Phi(x_n; \theta), y_n)$$

- The corresponding function we optimize is

$$g(\theta) = \mathbb{E}_{(x,y)\sim P}\ell(\Phi(x; \theta), y)$$

population or generalization error

CLASSIC STOCHASTIC APPROXIMATION (ROBBINS & MUNRO,'51)

- General problem: find the zeros of a function $h : \mathbb{R}^d \to \mathbb{R}^d$ from random observations at certain points.
  - Main motivation: $h = \nabla g$
- Original Robbins & Munro algorithm:

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

CLASSIC STOCHASTIC APPROXIMATION (ROBBINS & MUNRO,'51)

- General problem: find the zeros of a function $h : \mathbb{R}^d \to \mathbb{R}^d$ from random observations at certain points.
  - Main motivation: $h = \nabla g$
- Original Robbins & Munro algorithm:

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

- Many important questions:
  - Conditions for convergence
  - Choice of step size γ_n
  - Non-asymptotic behavior.

BIG PICTURE STOCHASTIC APPROXIMATION

- Suppose we want to use the previous algorithm to perform recursive mean estimation.
- Starting from $\theta_0 = 0$, we get data $x_n \in \mathbb{R}^d$:

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

BIG PICTURE STOCHASTIC APPROXIMATION

- Suppose we want to use the previous algorithm to perform recursive mean estimation.
- Starting from $\theta_0 = 0$, we get data $x_n \in \mathbb{R}^d$:

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

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

BIG PICTURE STOCHASTIC APPROXIMATION

- If x_n are iid with $\mathbb{E}x_n = x$, $\mathbb{E}\|x_n - x\|^2 = \sigma^2$:

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

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

BIG PICTURE STOCHASTIC APPROXIMATION

- If x_n are iid with $\mathbb{E}x_n = x\,, \mathbb{E}\|x_n - x\|_n^2 = \sigma^2$:

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

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

- Mean-Squared error has two contributions:
  - Forgetting initial conditions: $\prod_{k\leq n}(1-\gamma_k) \to 0 \;\text{ as } n\to\infty\,.$
  - Robustness to noise: $\sum_{i\leq n}\gamma_i^2\prod_{i<k\leq n}(1-\gamma_k)^2 \to 0 \;\text{ as } n\to\infty\,.$

FORGETTING INITIAL CONDITIONS

- If $\gamma_n \to 0$, $\log \prod_{k \leq n} (1 - \gamma_k) \simeq - \sum_{k \leq n} \gamma_k$, so γ_n should not decay too fast.

FORGETTING INITIAL CONDITIONS

- If $\gamma_n \to 0$, $\log \prod_{k \le n} (1 - \gamma_k) \simeq - \sum_{k \le n} \gamma_k$, so γ_n should not decay too fast.
- Example: $\gamma_n = C n^{-\alpha}$.
  - $\alpha = 1, \sum_{i<n} i^{-1} = \log(n) + C' + O(n^{-1})$,
  - $\alpha > 1, \sum_{i \le n} i^{-\alpha} = C' + O(n^{1-\alpha})$,
  - $\alpha \in (0,1)\,, \sum_{i \le n} i^{-\alpha} = C' n^{1-\alpha} + O(1)$.
- If $\alpha = 1$, squared norm converges with rate n^{-2C}.

NOISE TERM

- Suppose γ_n is non-increasing and smaller than $1/\mu$. Then

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

NOISE TERM

- Suppose γ_n is non-increasing and smaller than $1/\mu$. Then

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

- Example: $\gamma_n = Cn^{-\alpha}, \mu = 1$
  - $\alpha = 1$: convergence of noise term in $O(1/n)$ and forgetting of initial conditions in $O(n^{-2C})$: choose $C \geq 1/2$
  - $\alpha < 1$: noise term dominates.

CONVERGENCE OF STOCHASTIC OPTIMIZATION

- When g convex, show whether $\theta_n \to \theta_*$, or $g(\theta_n) \to g(\theta_*)$

CONVERGENCE OF STOCHASTIC OPTIMIZATION

- When g convex, show whether $\theta_n \to \theta_*$, or $g(\theta_n) \to g(\theta_*)$
- Since now these are random quantities, need to specify convergence criteria.
  - Convergence almost-surely: $P(g(\theta_n) \to g(\theta_*)) = 1$
  - Convergence in probability:

$$\forall\, \epsilon > 0,\ P(|g(\theta_n) - g(\theta_*)| \geq \epsilon) \to 0 \ .$$

  - Convergence in moments: $\mathbb{E}|g(\theta_n) - g(\theta_*)|^r \to 0$.

ROBBINS-MONRO ASYMPTOTIC NORMALITY

- Fabian, '68 shows asymptotic normality of Robbins Munro using traditional step-size $\gamma_n = Cn^{-1}$:

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

$$A = \nabla^2 g(\theta_*) \quad \Sigma = \mathbb{E}(\epsilon_n \epsilon_n^\top)$$

ROBBINS-MONRO ASYMPTOTIC NORMALITY

- Fabian, '68 shows asymptotic normality of Robbins Munro using traditional step-size $\gamma_n = Cn^{-1}$:

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

$$A = \nabla^2 g(\theta_*) \quad \Sigma = \mathbb{E}(\epsilon_n \epsilon_n^\top)$$

- It requires $2C\lambda_{\min}(A) \geq 1$ for convergence.
- C too small means no convergence, C too large: large variance.
- The conditioning of the problem impacts choice of step size.

ROBBINS-MONRO ASYMPTOTIC NORMALITY

- Fabian, '68 shows asymptotic normality of Robbins Munro using traditional step-size $\gamma_n = Cn^{-1}$:

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

$$A = \nabla^2 g(\theta_*) \quad \Sigma = \mathbb{E}(\epsilon_n \epsilon_n^\top)$$

- It requires $2C\lambda_{\min}(A) \geq 1$ for convergence.
- C too small means no convergence, C too large: large variance.
- The conditioning of the problem impacts choice of step size.
- Not ideal.

POLYAK–RUPPERT AVERAGING

- Robbins-Munro algorithm suffers from
  - Choice of step-size
  - Dependence on unknown conditioning of the problem.
- We can average the iterates to modify tradeoff bias-variance:

$$\overline{\theta}_n = \frac{1}{n} \sum_{k \leq n} \theta_k$$

- computed efficiently with recursion: $\overline{\theta}_n = (1 - \frac{1}{n})\overline{\theta}_{n-1} + n^{-1}\theta_n$
- It converges to a unique optimum

CESARO'S THEOREM

- Suppose $\theta_n \to \theta_*$ with convergence rate $\|\theta_n - \theta_*\| \leq \alpha_n$
- **Cesaro's Theorem:** $\overline{\theta}_n$ also converges to θ_*. Rate?

CESARO'S THEOREM

- Suppose $\theta_n \to \theta_*$ with convergence rate $\|\theta_n - \theta_*\| \leq \alpha_n$
- **Cesaro's Theorem:** $\overline{\theta}_n$ also converges to θ_*. Rate?

$$\|\overline{\theta}_n - \theta_*\| \leq \frac{1}{n}\sum_{k=1}^{n}\|\theta_k - \theta_*\| \leq \frac{1}{n}\sum_{k\leq n}\alpha_k$$

- Consequence: if $\sum_n \alpha_n < \infty$, rate is always $1/n$

PLETHORA OF CONVERGENCE RATES FOR STOCHASTIC OPT

- Global minimax rates of convergence for non-smooth problems (Nemirovsky et al'83, Agarwal et al'12):
  - Strongly convex: $O((\mu n)^{-1})$ with averaged stochastic gradient descent with $\gamma_n \propto (\mu n)^{-1}$.
  - Non-strongly convex: $O(n^{-1/2})$ with averaged stochastic gradient descent with $\gamma_n \propto n^{-1/2}$.
- Smooth problems: use $\gamma_n \propto n^{-1/2}$. with averaging to be adaptive to strong convexity [Bach and Moulines'11].

- What about constant step SGD?

LEAST SQUARES OPTIMIZATION

- Consider the problem

$$g(\theta) = \mathbb{E}[(Y - \langle X, \theta \rangle)^2], \theta \in \mathbb{R}^d.$$

LEAST SQUARES OPTIMIZATION

- Consider the problem

$$g(\theta) = \mathbb{E}[(Y - \langle X, \theta \rangle)^2] \, , \theta \in \mathbb{R}^d.$$

- For generic covariance $H = \mathbb{E}[XX^\top]$ (not necessarily pd) and *constant step-size* $\gamma = 1/(4R^2)$ (assuming $\|X\| \leq R$), averaged stochastic gradient descent satisfies

$$\mathbb{E}g(\overline{\theta}_{n-1}) - g(\theta_*) \leq \frac{C}{n} \ . \qquad \text{[Bach and Moulines'13]}$$

- Matches statistical lower-bound [Tsybakov,'03].
- More details in [Bach, Frejus'17].

STOCHASTIC GRADIENT DESCENT AND MARKOV CHAINS

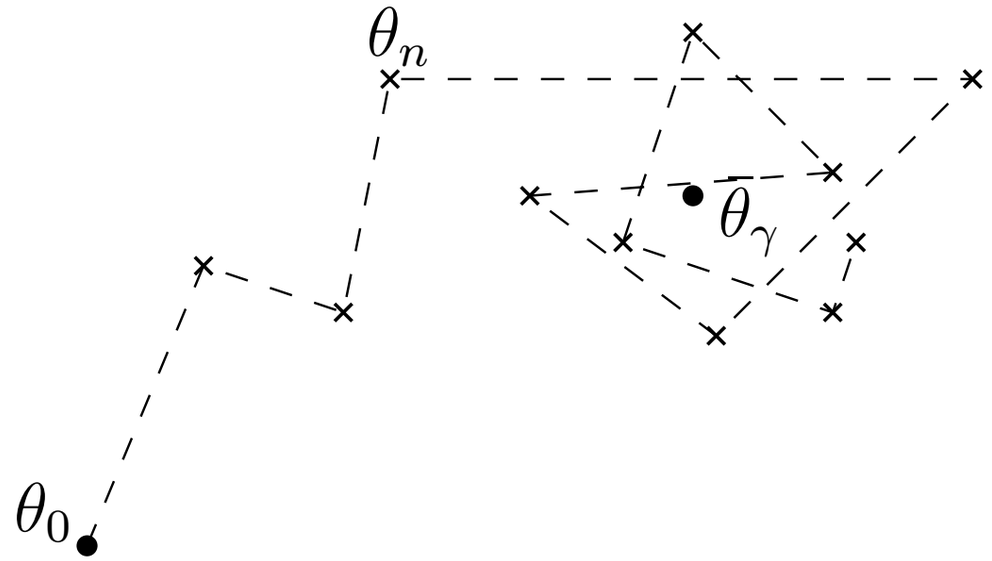
- The corresponding recursion of SGD in the Least Squares is

$$\theta_n = \theta_{n-1} - \gamma(\langle x_n, \theta_{n-1}\rangle - y_n)x_n \ .$$

- This defines a homogeneous Markov Chain:

$$\theta_n \,|\, \theta_{n-1} = (\mathbf{1} - \gamma x_n x_n^\top)\theta_{n-1} + \gamma y_n x_n \, .$$

- Convergence of the Chain to a stationary distribution π_γ.
- Its expectation is $\overline{\theta}_\gamma = \int \theta \pi_\gamma(d\theta)$



STOCHASTIC GRADIENT DESCENT AND MARKOV CHAINS

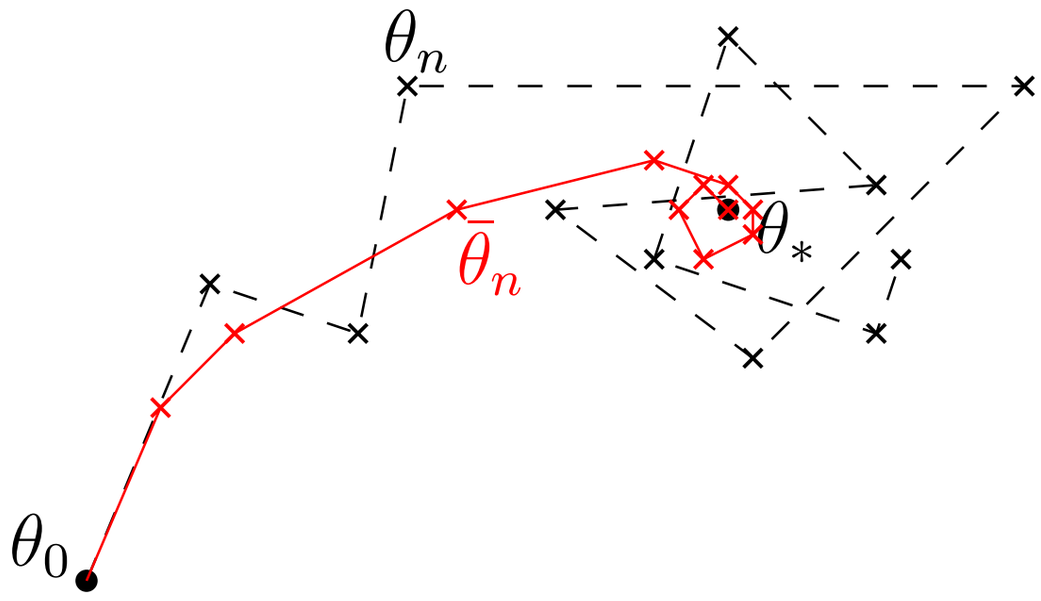
- For least squares, it turns out that
$$\overline{\theta}_\gamma = \theta_* = \mathbb{E}[XX^\top]^\dagger \mathbb{E}[X^\top Y].$$
- θ_n does NOT converge to θ_*, but oscillates around it, with oscillations of order $\sqrt{\gamma}$.

STOCHASTIC GRADIENT DESCENT AND MARKOV CHAINS

- For least squares, it turns out that

$$\overline{\theta}_\gamma = \theta_* = \mathbb{E}[XX^\top]^\dagger \mathbb{E}[X^\top Y].$$

- θ_n does NOT converge to θ_*, but oscillates around it, with oscillations of order $\sqrt{\gamma}$.
- But the averaged iterates do converge to this mean:
- Ergodic Theorem: $\overline{\theta}_n \to \overline{\theta}_\gamma$ at rate $O(1/n)$.



CONSTANT STEP SIZE BEYOND LEAST-SQUARES

- The general case (with constant step-size)

$$\theta_n = \theta_{n-1} - \gamma \nabla g_n(\theta_{n-1})$$

also defines a homogeneous Markov Chain.

CONSTANT STEP SIZE BEYOND LEAST-SQUARES

- The general case (with constant step-size)

$$\theta_n = \theta_{n-1} - \gamma \nabla g_n(\theta_{n-1})$$

  also defines a homogeneous Markov Chain.

- Its stationary distribution π_γ satisfies

$$\int \nabla g(\theta) \pi_\gamma(d\theta) = 0.$$

CONSTANT STEP SIZE BEYOND LEAST-SQUARES

- The general case (with constant step-size)

$$\theta_n = \theta_{n-1} - \gamma \nabla g_n(\theta_{n-1})$$

  also defines a homogeneous Markov Chain.

- Its stationary distribution π_γ satisfies

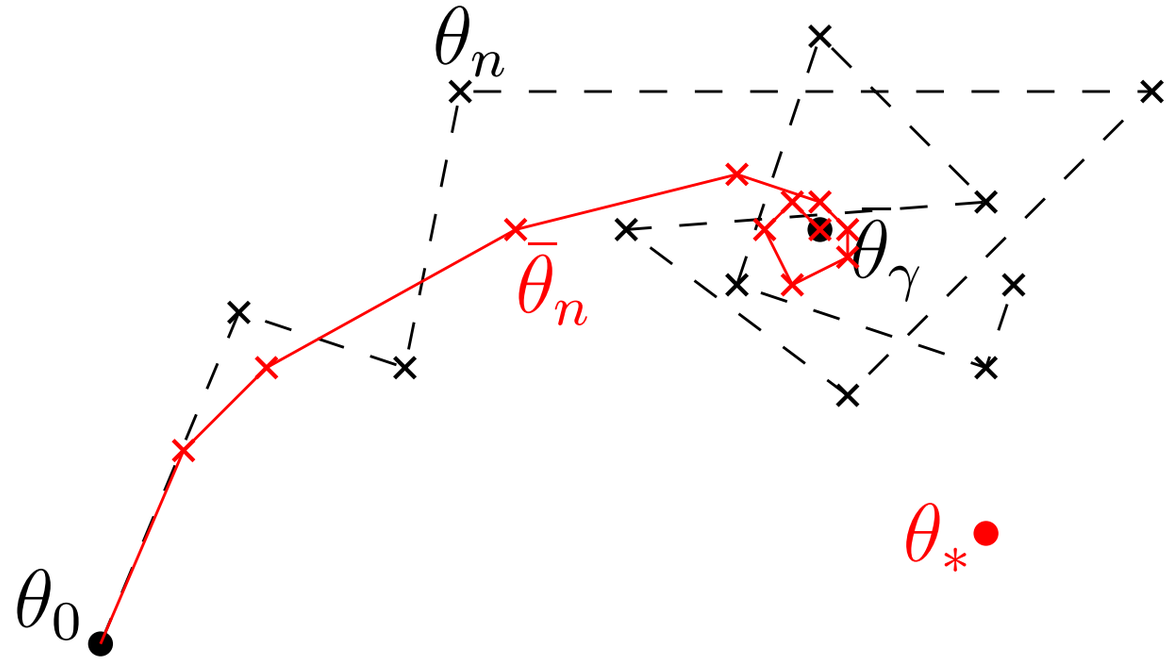
$$\int \nabla g(\theta) \pi_\gamma(d\theta) = 0.$$

- But in general, since ∇g is not linear,

$$\nabla g(\overline{\theta}_\gamma) = \nabla g \left(\int \theta \pi_\gamma(d\theta) \right) \neq \int \nabla g(\theta) \pi_\gamma(d\theta) = 0$$

CONSTANT STEP SIZE BEYOND LEAST SQUARES

- θ_n oscillates around the wrong stationary point $\overline{\theta}_\gamma \neq \theta_*$



- Moreover, $\|\theta_* - \overline{\theta}_\gamma\| = O(\gamma)$. [Bach'13]
- Fix?

BRIDGING THE GAP [DIEULEVEUT, DURMUS & BACH,'17]

- We consider again averaged iterates using constant step-size

$$\overline{\theta}_n = \frac{1}{n} \sum_{k \leq n} \theta_k$$

- Under appropriate conditions on the Markov chain $(\theta_k)_k$, we have a Central Limit Theorem for $(\theta_n)_n$, which means that $\overline{\theta}_n$ converges to $\int \theta \pi_\gamma(d\theta)$ at rate $O(n^{-1/2})$.
- Error decomposition:

$$\overline{\theta}_n - \theta_* = \underbrace{\overline{\theta}_n - \overline{\theta}_\gamma}_{\text{stochastic error}} + \underbrace{\overline{\theta}_\gamma - \theta_*}_{\text{deterministic error}}$$

BRIDGING THE GAP

- Under appropriate assumptions, we have that if g is strongly convex, then for small enough γ we have

$$\mathbb{E}(\overline{\theta}_k - \theta_*) = \frac{A(\theta_0, \gamma)}{k} + C\gamma + O(\gamma^2) + O(e^{-k\mu\gamma})$$

- Can we improve the convergence by exploiting this expansion?

BRIDGING THE GAP

- Under appropriate assumptions, we have that if g is strongly convex, then for small enough γ we have

$$\mathbb{E}(\overline{\theta}_k - \theta_*) = \frac{A(\theta_0, \gamma)}{k} + C\gamma + O(\gamma^2) + O(e^{-k\mu\gamma})$$

- Can we improve the convergence by exploiting this expansion?

- Richardson-Romberg extrapolation: consider two sequences $(\overline{\theta}_k^{(2\gamma)})_k$, $(\overline{\theta}_k^{(\gamma)})_k$. obtained by doubling the step-size. Then

$$\mathbb{E}(2\overline{\theta}_k^{(\gamma)} - \overline{\theta}_k^{(2\gamma)} - \theta_*) = \frac{2A(\theta_0, \gamma) - A(\theta_0, 2\gamma)}{k} + O(\gamma^2) + O(e^{-k\mu\gamma})$$

CONTINUOUS-TIME INTERPRETATION

- The error expansion in terms of the step-size can be seen as a discretization error of an underlying continuous (and stochastic) dynamics, given by the stochastic gradient flow

$$\dot{\theta}_t = -\nabla g(\theta_t) + \text{ noise } .$$

- How to deal with that kind of stochastic equation?
- Handle non-convex objectives as well?
- Relationship with other diffusion schemes?

STOCHASTIC DIFFERENTIAL EQUATIONS

- Let $T > 0$. An Ito stochastic differential equation on the interval [0,T] is of the form

$$dX_t = b(X_t, t)dt + \sigma(X_t, t)dW_t \ , \ \ X_0 = x_0 \ , \text{with}$$

$$X_t \in \mathbb{R}^d \ , b : \mathbb{R}^d \times [0, T] \to \mathbb{R}^d \ , \sigma : \mathbb{R}^d \times [0, T] \to \mathbb{R}^{d \times l}$$

$$W_t \text{: } l\text{-dimensional Brownian Motion.}$$

- A Brownian Motion (or Wiener process) W_t is such that:
  - $W_0 = 0$ a.s.
  - W_t has independent and Gaussian increments:
    $W_{t+u} - W_t \sim \mathcal{N}(0, u\mathbf{1})$,
    $W_{t+u} - W_t$ independent of $W_s, s < t$.
  - Sample paths are continuous with probability 1.

ITO CALCULUS

➤ This formalizes the intuition that SDEs are "noisy" ODEs:

$$\dot{\theta}_t = -\nabla g(\theta_t) + \text{ noise } .$$

$$\frac{dX_t}{dt} = b(X_t, t) + \sigma(X_t, t)\frac{dW_t}{dt}$$

➤ The SDE is associated with the stochastic integral equation

$$X_t - x_0 = \int_0^t b(X_s, s)ds + \underbrace{\int_0^t \sigma(X_s, s)dW_s}_{\text{Ito integral}}$$

$$\int_0^t F_s dW_s = \lim_{m\to\infty} \sum_{0=s_0<\cdots<s_m=t} F_{s_{i-1}}(W_{s_i} - W_{s_{i-1}}) .$$

DISCRETIZATION OF SDES

- Consider the SDE

$$dX_t = b(X_t, t)dt + \sigma(X_t, t)dW_t \ , \ X_0 = x_0 \, , \text{with}$$

- A stochastic discrete-time scheme $(x_k)_k$ approximates this SDE *in the weak sense* at order α if for every polynomial-growth function g $s.t.$ $|g(x)| \leq K(1 + |x|^\kappa)$ exists C>0 independent of δ such that for all $k = 0, 1, \ldots T/\delta$

$$|\mathbb{E}g(X_{k\delta}) - \mathbb{E}g(x_k)| \leq C\delta^\alpha \, .$$

- Weak approximations do not necessarily have similar sample paths, but rather their distribution.

DISCRETIZATIONS OF SDES: EULER MARUYAMA

- The Euler Maruyama method extends the Euler discretization of ODEs to SDEs.
- Consider as before the SDE

$$dX_t = b(X_t, t)dt + \sigma(X_t, t)dW_t \ , \ \ X_0 = x_0 \ , \text{with}$$

- Fix a time sampling interval $\delta > 0$ and define $x_k := X_{k\delta}$
- Consider the finite difference equation

$$x_{k+1} = x_k + \delta b(x_k, k\delta) + \sigma(x_k, k\delta)(W_{(k+1)\delta} - W_{k\delta}) \ .$$

- Since $W_{(k+1)\delta} - W_{k\delta} \sim \mathcal{N}(0, \delta \mathbf{1})$, this is equivalent to

$$x_{k+1} = x_k + \delta b(x_k, k\delta) + \sqrt{\delta}\sigma(x_k, k\delta)Z_k \ , \ Z_k \sim \mathcal{N}(0, \mathbf{1}) \ .$$

- This scheme is a first order weak approximation to the SDE.

STOCHASTIC MODIFIED EQUATIONS [LI ET AL.'17]

- Consider the Empirical Risk Minimization setup (no reg):

$$\min_{\theta \in \mathbb{R}^d} g(\theta) = \frac{1}{n} \sum_{i \leq n} g_i(\theta)$$

- Stochastic Gradient Descent with fixed step-size:

$$\theta_{k+1} = \theta_k - \gamma \nabla g_{m_k}(\theta_k)\,,$$
$$\{m_k\}\text{: iid uniform variables in } \{1, \dots, n\}.$$

- Goal: derive SDEs that can be seen as weak limits of stochastic gradient descent.
  - In deterministic systems, these are called *modified equations*.

STOCHASTIC MODIFIED EQUATIONS

- We decompose the stochastic gradient in terms of its mean:

$$V_k := \sqrt{\gamma}(\nabla g(\theta_k) - \nabla g_{m_k}(\theta_k))$$

- We have that

$$\theta_{k+1} - \theta_k = -\gamma \nabla g(\theta_k) + \sqrt{\gamma} V_k$$

- Conditioned on θ_k, V_k has mean 0 and covariance $\gamma\Sigma(\theta_k)$:

$$\Sigma(\theta) = \frac{1}{n} \sum_{i \leq n} (\nabla g(\theta) - \nabla g_i(\theta))(\nabla g(\theta) - \nabla g_i(\theta))^\top .$$

- Does this look similar to the Euler-Maruyama scheme for an appropriately chosen SDE?

STOCHASTIC MODIFIED EQUATIONS

- **Theorem [Li, Tai, E, '17]:** Let $T > 0$ and $\Sigma(\theta) \in \mathbb{R}^{d\times d}$ as before. Assuming g, g_k are sufficiently smooth, then

$$dX_t = -\nabla g(X_t)dt + (\gamma\Sigma(X_t))^{1/2}dW_t$$

is an order 1 weak approximation of the SGD.

$$dX_t = -\nabla(g(X_t) + \frac{\gamma}{4}\|\nabla g(X_t)\|^2)dt + (\gamma\Sigma(X_t))^{1/2}dW_t$$

is an order 2 weak approximation of the SGD.

- No convexity assumptions on g.
- How easy is to solve SDEs like above?

SGD DYNAMICS

- We obtained an SDE of the form

$$dX_t^{(\epsilon)} = b(X_t^{(\epsilon)})dt + \epsilon\sigma(X_t^{(\epsilon)})dW_t \ \text{, with } \epsilon \ll 1\,.$$

- Stochastic Asymptotic Expansions [Friedlin et al'12] consider asymptotic expansions wrt ϵ, assuming b and σ are smooth:

$$X_t^{(\epsilon)} = X_{0,t} + \epsilon X_{1,t} + \epsilon^2 X_{2,t} + \dots$$
$$b(X_t^{(\epsilon)}) = b(X_{0,t}) + \epsilon\nabla b(X_{0,t})X_{1,t} + O(\epsilon^2)$$
$$\sigma(X_t^{(\epsilon)}) = \sigma(X_{0,t}) + \epsilon\nabla\sigma(X_{0,t})X_{1,t} + O(\epsilon^2)$$

- Identifying terms with same order we obtain

$$dX_{0,t} = b(X_{0,t})dt$$
$$dX_{1,t} = \nabla b(X_{0,t})X_{1,t}dt + \sigma(X_{0,t})dW_t$$
$$\dots$$

SGD DYNAMICS

- Applying the Stochastic Asymptotic Expansion in the SGD case we obtain

$$X_t \sim \mathcal{N}(X_{0,t}, \gamma S_t) \ ,$$

$$\dot{X}_{0,t} = -\nabla g(X_{0,t}) \quad \dot{S}_t = -S_t H_t - H_t S_t + \Sigma_t$$

$$\text{with } H_t = \nabla^2 g(X_{0,t}), \Sigma_t = \Sigma(X_{0,t}) \ .$$

- This implies that the steady state of S_t satisfies $|S_\infty| \sim \dfrac{|\Sigma_\infty|}{|H_\infty|}$
- Two phases in SGD:
  - Descent regime dominated by gradient flow.
  - Fluctuating regime dominated by St.

LANGEVIN DYNAMICS

- Now suppose we modify SGD by adding a properly scaled isotropic Gaussian noise

- This is the Stochastic Gradient Langevin Dynamics.
- Its continuous limit is the Langevin diffusion

- The gibbs measure is the unique invariant distribution of this diffusion.
- For sufficiently large beta, the Gibbs distribution concentrates around the minimizers of g.

STOCHASTIC GRADIENT LANGEVIN DYNAMICS