

Gradient Estimation Methods for Discrete Event Systems Performance Measures Part II

Felisa J. Vázquez-Abad

Département d'informatique et recherche opérationnelle

Université de Montréal, Québec, CANADA

courriel: vazquez@IRO.UMontreal.CA

Department of Electronic and Electrical Engineering

The University of Melbourne

5th WODES, Ghent August 21 and 22, 2000.

Main Results: Pathwise Analysis

Definition: A random variable $h(\theta, \omega)$ on $(\Omega, \mathfrak{F}, \mathbb{P})$ is said to be **a.s. Lipschitz continuous in θ** if there is a random variable $K < \infty$ with $E[K] < \infty$ such that for each ω ,

$$\sup_{\theta \in \Theta: \theta + \Delta\theta \in \Theta} \|h(\theta + \Delta\theta, \omega) - h(\theta, \omega)\| \leq K(\omega) \Delta\theta,$$

If $h(\theta)$ is a.s. Lipschitz continuous in θ , then $E[h'(\theta)] = J'(\theta)$.

- Filtered Monte Carlo: for $i \leq n$, $J(\theta) = E[L(Z_1, \dots, Z_n) | \mathfrak{F}_i]$ is a r.v. which is measurable w.r.t. \mathfrak{F}_i . Call it $h_i(\theta, \omega)$. Then $J(\theta) = E[h_i], i \leq n$

$$J(\theta) = E \left[\int_{\Omega} h_i(\theta, \omega) \mathbb{P}(d\omega) \right]$$

Main Results: Pathwise Analysis

Definition: A random variable $h(\theta, \omega)$ on $(\Omega, \mathfrak{F}, \mathbb{P})$ is said to be **a.s. Lipschitz continuous in θ** if there is a random variable $K < \infty$ with $E[K] < \infty$ such that for each ω ,

$$\sup_{\theta \in \Theta: \theta + \Delta\theta \in \Theta} \|h(\theta + \Delta\theta, \omega) - h(\theta, \omega)\| \leq K(\omega) \Delta\theta,$$

If $h(\theta)$ is a.s. Lipschitz continuous in θ , then $E[h'(\theta)] = J'(\theta)$.

- **Filtered Monte Carlo:** for $i \leq n$, $J(\theta) = E[L(Z_1, \dots, Z_n) | \mathfrak{F}_i]$ is a r.v. which is measurable w.r.t. \mathfrak{F}_i . Call it $h_i(\theta, \omega)$. Then $J(\theta) = E[h_i], i \leq n$

$$J(\theta) = E \left[\int_{\Omega} h_i(\theta, \omega) \mathbb{P}(d\omega) \right]$$

- **Smoothing:** conditioning may integrate discontinuities: $h_i(\theta)$ a.s. Lipschitz continuous.

Main Results: Pathwise Analysis

Definition: A random variable $h(\theta, \omega)$ on $(\Omega, \mathfrak{F}, \mathbb{P})$ is said to be **a.s. Lipschitz continuous in θ** if there is a random variable $K < \infty$ with $E[K] < \infty$ such that for each ω ,

$$\sup_{\theta \in \Theta: \theta + \Delta\theta \in \Theta} \|h(\theta + \Delta\theta, \omega) - h(\theta, \omega)\| \leq K(\omega) \Delta\theta,$$

If $h(\theta)$ is a.s. Lipschitz continuous in θ , then $E[h'(\theta)] = J'(\theta)$.

- **Filtered Monte Carlo:** for $i \leq n$, $J(\theta) = E[L(Z_1, \dots, Z_n) | \mathfrak{F}_i]$ is a r.v. which is measurable w.r.t. \mathfrak{F}_i . Call it $h_i(\theta, \omega)$. Then $J(\theta) = E[h_i], i \leq n$

$$J(\theta) = E \left[\int_{\Omega} h_i(\theta, \omega) \mathbb{P}(d\omega) \right]$$

- **Smoothing:** conditioning may integrate discontinuities: $h_i(\theta)$ a.s. Lipschitz continuous.
- **Representations:** use alternative representations of $L(Z)$ as a function $h'(\theta\omega)$ (problem dependent: case by case).

Weak Differentiation

Definition: If $(X_n, n \in \mathbb{N})$ is a sequence of random variables on $(\Omega, \mathfrak{F}, \mathbb{P})$, and X is \mathfrak{F} -measurable,

- X_n converges to X almost surely, denoted by $X_n \rightarrow X$ a.s., if $\mathbb{P}\{\omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\} = 1$.
- X_n converges to X weakly, denoted by $X_n \xrightarrow{\mathcal{L}} X$, if for every bounded and continuous function $g : S \rightarrow \mathbb{R}$, $\mathbb{E}[g(X_n)] \rightarrow \mathbb{E}[g(X)]$.

Weak Differentiation

Definition: If $(X_n, n \in \mathbb{N})$ is a sequence of random variables on $(\Omega, \mathfrak{F}, \mathbb{P})$, and X is \mathfrak{F} -measurable,

- X_n **converges to X almost surely**, denoted by $X_n \rightarrow X$ **a.s.**, if $\mathbb{P}\{\omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\} = 1$.
- X_n **converges to X weakly**, denoted by $X_n \xrightarrow{\mathcal{L}} X$, if for every bounded and continuous function $g : S \rightarrow \mathbb{R}$, $\mathbb{E}[g(X_n)] \rightarrow \mathbb{E}[g(X)]$.

Definition: Let μ_θ be a family of probability measures defined on $(S, \mathcal{B}(S))$. The probability measure μ_θ is said to be **weakly differentiable** if **there exists a measure $\mu'_\theta : \mathcal{B}(S) \rightarrow \mathbb{R}$** such that for every $g \in C_b(S)$

$$\lim_{\Delta\theta \rightarrow 0} \frac{1}{\Delta\theta} \left(\int_S g(s) \mu_{\theta+\Delta\theta}(ds) - \int_S g(s) \mu_\theta(ds) \right) = \int_S g(s) \mu'_\theta(ds).$$

(and then $\frac{\partial}{\partial\theta} \mathbb{E}[g] = \int g(s) \mu'_\theta(ds)$).

REMARK: Note that $\mu_{\theta+\Delta\theta}(\cdot) - \mu_\theta(\cdot)$ is a signed measure. The weak limit (if it exists) is also a **finite signed measure with total mass zero**.

Examples of weak differentiation

EXAMPLE 1. $X \sim \exp(\theta)$. Score function :

$$\forall g \in C_b(\mathbb{R}) \frac{d}{d\theta} \int g(x) \theta e^{-\theta x} dx = \int g(x) \left(\frac{1}{\theta} - x \right) \theta e^{-\theta x} dx$$

Examples of weak differentiation

EXAMPLE 2. $X \sim \exp(\theta)$. Score function :

$$\forall g \in C_b(\mathbb{R}) \frac{d}{d\theta} \int g(x) \theta e^{-\theta x} dx = \int g(x) \left(\frac{1}{\theta} - x \right) \theta e^{-\theta x} dx$$

$$\mu'_\theta(dx) = (1/\theta - x)\theta e^{-\theta x} dx, \text{ not a density.}$$

Examples of weak differentiation

EXAMPLE 3. $X \sim \exp(\theta)$. Score function :

$$\forall g \in C_b(\mathbb{R}) \frac{d}{d\theta} \int g(x) \theta e^{-\theta x} dx = \int g(x) \left(\frac{1}{\theta} - x \right) \theta e^{-\theta x} dx$$

$$\mu'_\theta(dx) = (1/\theta - x)\theta e^{-\theta x} dx, \text{ not a density.}$$

EXAMPLE 4. $X \sim U[0, \theta]$ then

$$\begin{aligned} \forall g \in C_b([0, \theta]) \quad \frac{d}{d\theta} \int_0^\theta \frac{1}{\theta} g(x) dx &= -\frac{1}{\theta^2} \int g(x) dx + \frac{1}{\theta} g(\theta) \\ &= \frac{1}{\theta} \left[\int_{\mathbb{R}} \delta_\theta(x) g(x) dx - \int_0^\theta g(x) dx \right] \end{aligned}$$

and $\mu'(\theta)(A) = 1/\theta[\mathbf{1}_{\{\theta \in A\}} - \mu_\theta(A)]$.

Examples of weak differentiation

EXAMPLE 5. $X \sim \exp(\theta)$. Score function :

$$\forall g \in C_b(\mathbb{R}) \frac{d}{d\theta} \int g(x) \theta e^{-\theta x} dx = \int g(x) \left(\frac{1}{\theta} - x \right) \theta e^{-\theta x} dx$$

$$\mu'_\theta(dx) = (1/\theta - x)\theta e^{-\theta x} dx, \text{ not a density.}$$

EXAMPLE 6. $X \sim U[0, \theta]$ then

$$\begin{aligned} \forall g \in C_b([0, \theta]) \quad \frac{d}{d\theta} \int_0^\theta \frac{1}{\theta} g(x) dx &= -\frac{1}{\theta^2} \int g(x) dx + \frac{1}{\theta} g(\theta) \\ &= \frac{1}{\theta} \left[\int_{\mathbb{R}} \delta_\theta(x) g(x) dx - \int_0^\theta g(x) dx \right] \end{aligned}$$

and $\mu'(\theta)(A) = 1/\theta[\mathbf{1}_{\{\theta \in A\}} - \mu_\theta(A)]$.

The weak derivative μ'_θ is the difference of two measures:
a mass at θ and the original measure $U[0, \theta]$.

Weak Differentiation: Absolute Continuity

Definition: Consider the space $(S, \mathcal{B}(S))$.

- A measure μ is said to be **absolutely continuous** w.r.t. a measure ν , denoted by $\mu \ll \nu$, if for every $B \in \mathcal{B}(S)$ $\nu(B) = 0$ implies $\mu(B) = 0$.
- A family of measures $\mu_\theta, \theta \in \Theta$ is said to be **dominated by ν** if $\mu_\theta \ll \nu$, for all $\theta \in \Theta$.
- Two measures ν and μ are **orthogonal**, denoted $\mu \perp \nu$ if there is $A \in \mathcal{B}(S)$ with $\nu(A) = 0$ and $\mu(A^c) = 0$.

Weak Differentiation: Absolute Continuity

Definition: Consider the space $(S, \mathcal{B}(S))$.

- A measure μ is said to be **absolutely continuous** w.r.t. a measure ν , denoted by $\mu \ll \nu$, if for every $B \in \mathcal{B}(S)$ $\nu(B) = 0$ implies $\mu(B) = 0$.
- A family of measures $\mu_\theta, \theta \in \Theta$ is said to be **dominated by ν** if $\mu_\theta \ll \nu$, for all $\theta \in \Theta$.
- Two measures ν and μ are **orthogonal**, denoted $\mu \perp \nu$ if there is $A \in \mathcal{B}(S)$ with $\nu(A) = 0$ and $\mu(A^c) = 0$.

Theorem 1 (Radon-Nikodym) Two σ -finite measures on $\mathcal{B}(S)$ satisfy $\mu \ll \nu$ if and only if there exists a nonnegative measurable function on $(S, \mathcal{B}(S))$ denoted by $\left[\frac{d\mu}{d\nu} \right] : S \rightarrow \mathbb{R}$ such that for every $B \in \mathcal{B}(S)$

$$\mu(B) = \int_B \left[\frac{d\mu}{d\nu} \right] (s) \nu(ds).$$

The function $\left[\frac{d\mu}{d\nu} \right]$ is called the **Radon-Nikodym derivative** and sometimes it is also called the “ ν -density” of μ .

Weak Differentiation: Absolute Continuity

Dominated family $\mu_\theta \ll \nu$ of **probability measures**, then:

$$\mu(B) = \int_B \left[\frac{d\mu}{d\nu} \right] (s) \nu(ds) = \int_B f_\theta(s) \nu(ds).$$

• **Likelihood Ratio:** Underlying space (Ω, \mathfrak{F}) , $X(\theta) \sim \mu_\theta$ and $Y \sim \nu$.
Define the nonnegative $\sigma(Y)$ -measurable random variable:

$$L(\theta, \omega) = f_\theta(Y(\omega))$$

then for all $g \in C_b(S)$ $\mathbf{E}[g(Y)L(\theta, Y)] = \mathbf{E}[g(X(\theta))]$. Change of measure.

Weak Differentiation: Absolute Continuity

Dominated family $\mu_\theta \ll \nu$ of **probability measures**, then:

$$\mu(B) = \int_B \left[\frac{d\mu}{d\nu} \right] (s) \nu(ds) = \int_B f_\theta(s) \nu(ds).$$

• **Likelihood Ratio:** Underlying space (Ω, \mathfrak{F}) , $X(\theta) \sim \mu_\theta$ and $Y \sim \nu$. Define the nonnegative $\sigma(Y)$ -measurable random variable:

$$L(\theta, \omega) = f_\theta(Y(\omega))$$

then for all $g \in C_b(S)$ $\mathbf{E}[g(Y)L(\theta, Y)] = \mathbf{E}[g(X(\theta))]$. Change of measure.

• **Score function method:**

$$\begin{aligned} \frac{d}{d\theta} \int g(s) \mu_\theta(ds) &= \int g(s) \frac{d}{d\theta} f_\theta(s) \nu(ds) = \int g(s) \left(\frac{\frac{d}{d\theta} f_\theta(s)}{f_\theta(s)} \right) f_\theta(s) \nu(ds) \\ &= \int g(s) \frac{d}{d\theta} \ln[f_\theta(s)] f_\theta(s) \nu(ds) = \int g(s) S(\theta, s) \mu_\theta(ds) \end{aligned}$$

Score Function for DES

$\{Z_n(\theta)\}$ a homogeneous Markov process in a general state space $S \subset \mathbb{R}^d$. Transitions **dominated by lebesgue measure** (densities):

$$\forall z \in S, \quad P_\theta(z, A) = \mathbb{P}(Z_{n+1} \in A | Z_n = z) \sim f_\theta(z, x)$$

Chain rule for differentiation \Rightarrow

$$\begin{aligned} J'(\theta) &= \frac{d}{d\theta} \int g(s_1, \dots, s_T) \prod_{i=1}^T P_\theta(ds_i, s_{i-1}) \\ &= \sum_{j=1}^T \int g(s_1, \dots, s_T) \frac{d}{d\theta} \ln(f_\theta(s_j, s_{j-1})) \times \\ &\quad \prod_{i=j+1}^T P_\theta(ds_i, s_{i-1}) P_\theta(ds_j, s_{j-1}) \prod_{i=1}^{j-1} P_\theta(ds_i, s_{i-1}), \end{aligned}$$

for any $g \in C_b(S)$. **Gradient estimators:**

$$\frac{d}{d\theta} \mathbb{E}[L(Z_1, \dots, Z_T)] = \sum_{j=1}^n \mathbb{E} \left[L(Z_1, \dots, Z_T) \frac{d}{d\theta} \ln(f_\theta(Z_j, Z_{j-1})) \right]$$

Weak Differentiation

EXAMPLE 7. $X \sim U[0, \theta]$ then

- **Weak derivative:**

$$\begin{aligned} \forall g \in C_b([0, \theta]) \quad \frac{d}{d\theta} \int_0^\theta \frac{1}{\theta} g(x) dx &= -\frac{1}{\theta^2} \int_0^\theta g(x) dx + \frac{1}{\theta} g(\theta) \\ &= \frac{1}{\theta} \left[\int_{\mathbb{R}} \delta_\theta(x) g(x) dx - \frac{1}{\theta} \int_0^\theta g(x) dx \right] \end{aligned}$$

and $\mu'(\theta)(A) = 1/\theta[\mathbf{1}_{\{\theta \in A\}} - \mu_\theta(A)]$.

- **Score function:** write down the (usual) density :

$$f_\theta(x) = \begin{cases} \frac{1}{\theta} & x \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

What is $\frac{d}{d\theta} \ln f_\theta(x)$????

Score function is not applicable: support depends on $\theta \Rightarrow$ unbounded derivatives of densities.

Weak Derivatives

Theorem 2 (Hahn-Jordan Decomposition) For any finite signed measure ν on $(S, \mathcal{B}(S))$ there exist disjoint sets S^+, S^- such that $S^+ \cup S^- = S$, with the property that

$$A \subset S^+ \Rightarrow \nu(A) > 0 \quad \text{and} \quad A \subset S^- \Rightarrow \nu(A) < 0$$

Definition: The **positive and negative** parts of ν are:

$$\nu_{\theta}^{\pm}(B) = \nu_{\theta}(B \cap S^{\pm}) \quad \forall B \subset \mathcal{B}(S)$$

so that $\nu(B) = \nu^+(B) + \nu^-(B)$.

Weak Derivatives

Theorem 3 (Hahn-Jordan Decomposition) For any finite signed measure ν on $(S, \mathcal{B}(S))$ there exist disjoint sets S^+, S^- such that $S^+ \cup S^- = S$, with the property that

$$A \subset S^+ \Rightarrow \nu(A) > 0 \quad \text{and} \quad A \subset S^- \Rightarrow \nu(A) < 0$$

Definition: The **positive and negative** parts of ν are:

$$\nu_\theta^\pm(B) = \nu_\theta(B \cap S^\pm) \quad \forall B \subset \mathcal{B}(S)$$

so that $\nu(B) = \nu^+(B) + \nu^-(B)$.

- In particular, if μ_θ is weakly differentiable then $\mu'_\theta(S) = 0$,
- Hahn-Jordan: $c_\theta = (\mu')_\theta^+(S) = -(\mu')_\theta^-(S)$,
- **Normalising:** Let $\mu_\theta^\pm(B) = (\mu')_\theta^\pm(B)/c_\theta \quad \forall B \subset \mathcal{B}(S)$
- Then μ_θ^\pm are **probability measures**

Using these probabilities, the derivative of an expectation can be expressed as the difference of two expectations.

Weak Derivatives

Definition: Let μ_θ be a family of measures on $(S, \mathcal{B}(S))$. A triple $(c_\theta, \mu_\theta^+, \mu_\theta^-)$ consisting of

- a positive constant c_θ ,
- two probability measures μ_θ^\pm

and is called a **weak derivative of μ_θ** if

- μ_θ is **weakly differentiable** at each θ , and
- For all $g \in C_b(S)$

$$\int_S g(s) \mu'_\theta(ds) = c_\theta \left(\int_S g(s) \mu_\theta^+(ds) - \int_S g(s) \mu_\theta^-(ds) \right)$$

If the left-hand side equals zero for all g , we define the weak derivative of μ_θ as $(1, \mu_\theta, \mu_\theta)$.

EXAMPLE 8. For $X \sim U[0, \theta]$, $c_\theta = \frac{1}{\theta}$, μ_θ^+ is the mass density at θ , and $\mu_\theta^- = \mu_\theta \sim U[0, \theta]$.

Overview: examples

EXAMPLE 9. $X \sim \exp(\theta)$, domination $\nu \sim$ lebesgue measure.
 $f_\theta(x) = \theta e^{-\theta x}$.

$$\frac{d}{d\theta} f_\theta(x) = e^{-\theta x} - \theta x e^{-\theta x} \quad (> 0 \text{ iff } x > 1/\theta)$$

- Use **Hahn-Jordan**:

$$c_\theta = \int_0^{1/\theta} (1 - \theta x) e^{-\theta x} dx = e\theta$$

$$f_\theta^+(x) = (1 - \theta x) e^{1-\theta x} 1_{\{0 < x \leq 1/\theta\}}$$

$$f_\theta^-(x) = -(1 - \theta x) e^{1-\theta x} 1_{\{x > 1/\theta\}}$$

Overview: examples

EXAMPLE 10. $X \sim \exp(\theta)$, domination $\nu \sim$ lebesgue measure.
 $f_\theta(x) = \theta e^{-\theta x}$.

$$\frac{d}{d\theta} f_\theta(x) = e^{-\theta x} - \theta x e^{-\theta x} \quad (> 0 \text{ iff } x > 1/\theta)$$

- Use **Hahn-Jordan**:

$$c_\theta = \int_0^{1/\theta} (1 - \theta x) e^{-\theta x} dx = e\theta$$

$$f_\theta^+(x) = (1 - \theta x) e^{1-\theta x} 1_{\{0 < x \leq 1/\theta\}}$$

$$f_\theta^-(x) = -(1 - \theta x) e^{1-\theta x} 1_{\{x > 1/\theta\}}$$

- **Simulation:** Generate **two random variables**, X^\pm with respective densities $f_\theta^\pm(\cdot)$ and build the estimator $c_\theta[L(X^+) - L(X^-)]$

Overview: examples

EXAMPLE 11. $X \sim \exp(\theta)$, domination $\nu \sim$ lebesgue measure.
 $f_\theta(x) = \theta e^{-\theta x}$.

$$\frac{d}{d\theta} f_\theta(x) = e^{-\theta x} - \theta x e^{-\theta x} \quad (> 0 \text{ iff } x > 1/\theta)$$

- Use **Hahn-Jordan**:

$$c_\theta = \int_0^{1/\theta} (1 - \theta x) e^{-\theta x} dx = e\theta$$

$$f_\theta^+(x) = (1 - \theta x) e^{1-\theta x} 1_{\{0 < x \leq 1/\theta\}}$$

$$f_\theta^-(x) = -(1 - \theta x) e^{1-\theta x} 1_{\{x > 1/\theta\}}$$

- **Simulation:** Generate **two random variables**, X^\pm with respective densities $f_\theta^\pm(\cdot)$ and build the estimator $c_\theta[L(X^+) - L(X^-)]$
- **Score function:** Uses estimator $L(X)(1/\theta - X)$,
- **IPA:** Uses estimator $h'(\theta)(X/\theta)$ (chain rule), if unbiased.
