Convergence of Adaptive Importance Samplers for Unbounded Parametric Families

Carlos A.C.C. Perello[1] Deniz Akyildiz[1]

[1]Department of Mathematics, Imperial College London

July 18, 2023

Contents

Contents

Importance Sampling

Set-up:

Importance Sampling

Set-up:

- Unknown target distribution π - can only evaluate its unnormalised density $\Pi(x)$ pointwise

Importance Sampling

Set-up:

- Unknown target distribution π - can only evaluate its unnormalised density $\Pi(x)$ pointwise
- Known proposal distribution q

Importance Sampling

Set-up:

- Unknown target distribution π - can only evaluate its unnormalised density $\Pi(x)$ pointwise
- Known proposal distribution q
- Known bounded function $\phi : X \to \mathbb{R}$

Importance Sampling

Set-up:

- Unknown target distribution π - can only evaluate its unnormalised density $\Pi(x)$ pointwise
- Known proposal distribution q
- Known bounded function $\phi : X \to \mathbb{R}$
- $\text{Supp}(\phi) \subseteq \text{Supp}(\pi) \subseteq \text{Supp}(q)$

Importance Sampling

Set-up:

- Unknown target distribution π - can only evaluate its unnormalised density $\Pi(x)$ pointwise
- Known proposal distribution q
- Known bounded function $\phi : X \to \mathbb{R}$
- $\text{Supp}(\phi) \subseteq \text{Supp}(\pi) \subseteq \text{Supp}(q)$

We want to estimate $(\phi, \pi) = \int_X \phi(x)\pi(x)dx$. Let $Z = \int_{\mathbb{R}^d} \Pi(x)dx$. Then:

$$(\phi, \pi) = \int_X \phi(x)\pi(x)dx$$

Importance Sampling

Set-up:

- Unknown target distribution π - can only evaluate its unnormalised density $\Pi(x)$ pointwise
- Known proposal distribution q
- Known bounded function $\phi : X \to \mathbb{R}$
- $\text{Supp}(\phi) \subseteq \text{Supp}(\pi) \subseteq \text{Supp}(q)$

We want to estimate $(\phi, \pi) = \int_X \phi(x)\pi(x)dx$. Let $Z = \int_{\mathbb{R}^d} \Pi(x)dx$. Then:

$$(\phi, \pi) = \int_X \phi(x)\pi(x)dx = \frac{\int_X \phi(x)\Pi(x)dx}{\int_{\mathbb{R}^d} \Pi(x)dx}$$

Importance Sampling

Set-up:

- Unknown target distribution π - can only evaluate its unnormalised density $\Pi(x)$ pointwise
- Known proposal distribution q
- Known bounded function $\phi : X \to \mathbb{R}$
- $\mathrm{Supp}(\phi) \subseteq \mathrm{Supp}(\pi) \subseteq \mathrm{Supp}(q)$

We want to estimate $(\phi, \pi) = \int_X \phi(x)\pi(x)dx$. Let $Z = \int_{\mathbb{R}^d} \Pi(x)dx$. Then:

$$(\phi, \pi) = \int_X \phi(x)\pi(x)dx = \frac{\int_X \phi(x)\Pi(x)dx}{\int_{\mathbb{R}^d} \Pi(x)dx} = \frac{\int_X \phi(x)\frac{\Pi(x)}{q(x)}q(x)dx}{\int_{\mathbb{R}^d} \frac{\Pi(x)}{q(x)}q(x)dx}$$

Importance Sampling

Define $W(x) := \frac{\Pi(x)}{q(x)}$. Using this:

$$(\phi, \pi) = \frac{\int_X \phi(x) W(x) q(x) dx}{\int_{\mathbb{R}^d} W(x) q(x) dx}$$

Importance Sampling

Define $W(x) := \frac{\Pi(x)}{q(x)}$. Using this:

$$(\phi, \pi) = \frac{\int_X \phi(x)W(x)q(x)dx}{\int_{\mathbb{R}^d} W(x)q(x)dx} \approx \sum_{i=1}^{N} \phi(x_i) \overbrace{\frac{W(x_i)}{\sum_{j=1}^{N} W(x_j)}}^{w(x_i)}$$

Importance Sampling

Define $W(x) := \frac{\Pi(x)}{q(x)}$. Using this:

$$(\phi, \pi) = \frac{\int_X \phi(x) W(x) q(x) dx}{\int_{\mathbb{R}^d} W(x) q(x) dx} \approx \sum_{i=1}^{N} \phi(x_i) \overbrace{\frac{W(x_i)}{\sum_{j=1}^N W(x_j)}}^{w(x_i)} = \sum_{i=1}^{N} \phi(x_i) w(x_i)$$

Setting $\tilde{\pi}^N(dx) = \sum_{i=1}^N w(x_i) \delta_{x_i}(dx)$ gives $(\phi, \pi) \approx (\phi, \tilde{\pi}^N)$

Definition

We call $\tilde{\pi}^N$ the *approximation/empirical measure* and N the *number of points/atoms* used to construct it. $(\phi, \tilde{\pi}^N)$ is the *Self-Normalised Importance Sampling (SNIS) estimator*.

Importance Sampling

Theorem 2.6. (Akyildiz and Míguez 2021)

If $(W^2, q) < \infty$ then:

$$\mathbb{E}[|(\phi, \pi) - (\phi, \tilde{\pi}^N)|^2] \leq \frac{4\|\phi\|_\infty^2 \rho}{N}$$

Where $\rho := \mathbb{E}_q\left[\frac{\pi^2(X)}{q^2(X)}\right]$. The same bound holds for π^N.

Importance Sampling

Theorem 2.6. (Akyildiz and Míguez 2021)

If $(W^2, q) < \infty$ then:

$$\mathbb{E}[|(\phi, \pi) - (\phi, \tilde{\pi}^N)|^2] \leq \frac{4\|\phi\|_\infty^2 \rho}{N}$$

Where $\rho := \mathbb{E}_q\left[\frac{\pi^2(X)}{q^2(X)}\right]$. The same bound holds for π^N.

Remark

$\rho = D_\chi(\pi\|q) + 1$. We call ρ the *Second Moment Error Metric (SMEM)*.

Optimised Adaptive Importance Sampling

Suppose the proposal is $q = q_\theta$. Then:

Optimised Adaptive Importance Sampling

Suppose the proposal is $q = q_\theta$. Then:

$$\rho = \rho(\theta), \quad \rho(\theta) = \mathbb{E}_{q_\theta}\left[\frac{\pi^2(X)}{q_\theta^2(X)}\right]$$

$\rho(\theta)$ is convex for q_θ in the exponential family (Akyildiz and Míguez 2021).

Optimised Adaptive Importance Sampling

Suppose the proposal is $q = q_\theta$. Then:

$$\rho = \rho(\theta), \quad \rho(\theta) = \mathbb{E}_{q_\theta}\left[\frac{\pi^2(X)}{q_\theta^2(X)}\right]$$

$\rho(\theta)$ is convex for q_θ in the exponential family (Akyildiz and Míguez 2021). For such q_θ, we can optimise θ and minimise $\rho(\theta)$. An algorithm that minimises $\rho(\theta)$ to adapt the proposal is an *OAIS algorithm*.

Optimised Adaptive Importance Sampling

Suppose the proposal is $q = q_\theta$. Then:

$$\rho = \rho(\theta), \quad \rho(\theta) = \mathbb{E}_{q_\theta}\left[\frac{\pi^2(X)}{q_\theta^2(X)}\right]$$

$\rho(\theta)$ is convex for q_θ in the exponential family (Akyildiz and Míguez 2021). For such q_θ, we can optimise θ and minimise $\rho(\theta)$. An algorithm that minimises $\rho(\theta)$ to adapt the proposal is an *OAIS algorithm*.

Remark

Oftentimes one may not be able to evaluate $\rho(\theta)$ either, but only an unnormalised version. We denote this version as $R(\theta)$.

Optimised Adaptive Importance Sampling

Algorithm 1 General OAIS algorithm

Choose a proposal q_θ with initial parameter θ_0 and a number of particles N.
for $t \geq 0$ **do**
    Sample $(x_t^{(i)})_{i=1}^N \sim q_{\theta_t}$
    Construct $\tilde{\pi}_t^N(dx) = \sum_{i=1}^N w(x_t^{(i)})\delta_{x_t^{(i)}}(dx)$
    Report $(\phi, \tilde{\pi}_t^N)$ and q_{θ_t}
    Compute the updated parameter θ_{t+1}[1]
end for

[1] Ideally using $(x_t^{(i)})_{i=1}^N$

Optimised Adaptive Importance Sampling

Algorithm 1 General OAIS algorithm

1: Choose a proposal q_θ with initial parameter θ_0 and a number of particles N.
2: **for** $t \geq 0$ **do**
3: Sample $(x_t^{(i)})_{i=1}^N \sim q_{\theta_t}$
4: Construct $\tilde{\pi}_t^N(dx) = \sum_{i=1}^N w(x_t^{(i)})\delta_{x_t^{(i)}}(dx)$
5: Report $(\phi, \tilde{\pi}_t^N)$ and q_{θ_t}
6: Compute the updated parameter θ_{t+1}[1]
7: **end for**

How does one update the parameters?

[1] Ideally using $(x_t^{(i)})_{i=1}^N$

Algorithm 1 General OAIS algorithm

1: Choose a proposal q_θ with initial parameter θ_0 and a number of particles N.
2: **for** $t \geq 0$ **do**
3: Sample $(x_t^{(i)})_{i=1}^N \sim q_{\theta_t}$
4: Construct $\tilde{\pi}_t^N(dx) = \sum_{i=1}^N w(x_t^{(i)})\delta_{x_t^{(i)}}(dx)$
5: Report $(\phi, \tilde{\pi}_t^N)$ and q_{θ_t}
6: Compute the updated parameter θ_{t+1}[1]
7: **end for**

How does one update the parameters? Minimising $\rho(\theta)$ using a gradient estimator $g(\theta) \rightsquigarrow$ Optimisation

[1] Ideally using $(x_t^{(i)})_{i=1}^N$

Optimise Adaptive Importance Sampling

Assumption 3.1.

$\rho(\theta)$ is convex and L-smooth w.r.t. the norm $\|\cdot\|_\Theta$, the parameter space's 2-norm.

Assumption 3.2.

The gradient of $\rho(\theta)$ is bounded: $\exists M > 0$ s.t. $\forall \theta \in \Theta$, $\|\nabla \rho(\theta)\|_2 \leq M$.

Optimised Adaptive Importance Sampling

Given Assumptions 3.1 and 3.2, (Akyildiz and Míguez 2021) prove that, after T iterations:

$$\mathbb{E}[|(\phi,\,\pi)-(\phi,\,\tilde{\pi}_T^N)|^2] \leq \frac{C_1}{\sqrt{T}N} + \frac{C_2}{\sqrt{T}N^2} + \frac{C_3}{\sqrt{T}N}(2+\log T) + \frac{C_4}{N}$$

However, they also assume that the parameter space Θ is compact.

Optimised Adaptive Importance Sampling

Given Assumptions 3.1 and 3.2, (Akyildiz and Míguez 2021) prove that, after T iterations:

$$\mathbb{E}[|(\phi,\, \pi) - (\phi,\, \tilde{\pi}_T^N)|^2] \leq \frac{C_1}{\sqrt{T}N} + \frac{C_2}{\sqrt{T}N^2} + \frac{C_3}{\sqrt{T}N}(2 + \log T) + \frac{C_4}{N}$$

However, they also assume that the parameter space Θ is compact. Furthermore, no numerical simulations are provided, and therefore the algorithm remained empirically unverified.

Optimised Adaptive Importance Sampling

Given Assumptions 3.1 and 3.2, (Akyildiz and Míguez 2021) prove that, after T iterations:

$$\mathbb{E}[|(\phi, \pi) - (\phi, \tilde{\pi}_T^N)|^2] \leq \frac{C_1}{\sqrt{T}N} + \frac{C_2}{\sqrt{T}N^2} + \frac{C_3}{\sqrt{T}N}(2 + \log T) + \frac{C_4}{N}$$

However, they also assume that the parameter space Θ is compact. Furthermore, no numerical simulations are provided, and therefore the algorithm remained empirically unverified.

Definition

If an OAIS algorithm has rate $\mathcal{O}(f(T)/N + 1/N)$ where $f(T) \to 0$ as $T \to \infty$, we call $\mathcal{O}(f(T))$ its *adaptive rate*.

Optimised Adaptive Importance Sampling

Given Assumptions 3.1 and 3.2, (Akyildiz and Míguez 2021) prove that, after T iterations:

$$\mathbb{E}[|(\phi,\, \pi) - (\phi,\, \tilde{\pi}_T^N)|^2] \leq \frac{C_1}{\sqrt{T}N} + \frac{C_2}{\sqrt{T}N^2} + \frac{C_3}{\sqrt{T}N}(2 + \log T) + \frac{C_4}{N}$$

However, they also assume that the parameter space Θ is compact. Furthermore, no numerical simulations are provided, and therefore the algorithm remained empirically unverified.

Definition

If an OAIS algorithm has rate $\mathcal{O}(f(T)/N + 1/N)$ where $f(T) \to 0$ as $T \to \infty$, we call $\mathcal{O}(f(T))$ its *adaptive rate*.

Goal: Obtain OAIS (adaptive) convergence rates without constraining Θ.

Contents

SG & USG OAIS

Given that Assumptions 3.1 & 3.2 hold on $\rho(\theta)$ in addition to mild assumptions on g, then using $t_k = \frac{C}{\sqrt{k+1}}$ after T iterations of SG OAIS:

SG & USG OAIS

Given that Assumptions 3.1 & 3.2 hold on $\rho(\theta)$ in addition to mild assumptions on g, then using $t_k = \frac{C}{\sqrt{k+1}}$ after T iterations of SG OAIS:

Theorem 3.2.

$$\mathbb{E}[|(\phi, \pi) - (\phi, \tilde{\pi}_T^N)|^2] \leq \frac{K_1 \mathbb{E}[\|\theta_0 - \theta^*\|_\Theta^2]}{N\sqrt{T+1}} + \frac{K_2 \log(T+1)}{N\sqrt{T+1}} + \frac{4\|\phi\|_\infty^2 \rho(\theta^*)}{N}$$

If USG is run instead with $R(\theta)$ and gradient estimator G satisfying Assumption 2.2 with S^2 instead of σ^2, we have the bound:

$$\mathbb{E}[|(\phi, \pi) - (\phi, \tilde{\pi}_T^N)|^2] \leq \frac{K_1 \mathbb{E}[\|\theta_0 - \theta^*\|_\Theta^2]}{N\sqrt{T+1}} + \frac{K_2' \log(T+1)}{Z^2 N\sqrt{T+1}} + \frac{4\|\phi\|_\infty^2 R(\theta^*)}{Z^2 N}$$

Where $K_1 = K_1(\phi, C)$, $K_2 = K_2(\phi, C, \sigma)$ and $K_2' = K_2'(\phi, C, S)$.

SG & USG OAIS

Given that Assumptions 3.1 & 3.2 hold on $\rho(\theta)$ in addition to mild assumptions on g, then using $t_k = \frac{C}{\sqrt{k+1}}$ after T iterations of SG OAIS:

Theorem 3.2.

$$\mathbb{E}[|(\phi, \pi) - (\phi, \tilde{\pi}_T^N)|^2] \leq \frac{K_1 \mathbb{E}[\|\theta_0 - \theta^*\|_\Theta^2]}{N\sqrt{T+1}} + \frac{K_2 \log(T+1)}{N\sqrt{T+1}} + \frac{4\|\phi\|_\infty^2 \rho(\theta^*)}{N}$$

If USG is run instead with $R(\theta)$ and gradient estimator G satisfying Assumption 2.2 with S^2 instead of σ^2, we have the bound:

$$\mathbb{E}[|(\phi, \pi) - (\phi, \tilde{\pi}_T^N)|^2] \leq \frac{K_1 \mathbb{E}[\|\theta_0 - \theta^*\|_\Theta^2]}{N\sqrt{T+1}} + \frac{K_2' \log(T+1)}{Z^2 N\sqrt{T+1}} + \frac{4\|\phi\|_\infty^2 R(\theta^*)}{Z^2 N}$$

Where $K_1 = K_1(\phi, C)$, $K_2 = K_2(\phi, C, \sigma)$ and $K_2' = K_2'(\phi, C, S)$.

Theorem 3.2 is novel in the IS/OAIS setting.

Theorem 3.2 is novel in the IS/OAIS setting.

The result was proven using a last-iterate SGD result which does not put any restrictions on the domain of f (Orabona 2020).

Adam OAIS

We will see later that SG OAIS is numerically unstable, so we adapted our results to adaptive optimisers.

Adam OAIS

We will see later that SG OAIS is numerically unstable, so we adapted our results to adaptive optimisers. A few assumptions are needed to give meaningful results about their convergence (Défossez et al. 2020).

Adam OAIS

We will see later that SG OAIS is numerically unstable, so we adapted our results to adaptive optimisers. A few assumptions are needed to give meaningful results about their convergence (Défossez et al. 2020).

Assumption 3.4.

Assume that the ℓ_∞ norm of the gradient estimators of ρ and R, g and G respectively, are almost surely-bounded; that is $\exists R_1,\ R_2 \geq \sqrt{\varepsilon}$, such that $\forall x \in \mathbb{R}^d$:

$$\|g(x)\|_\infty \leq R_1 - \sqrt{\varepsilon} \quad \text{a.s.},$$
$$\|G(x)\|_\infty \leq R_2 - \sqrt{\varepsilon} \quad \text{a.s.}$$

Adam OAIS

We will see later that SG OAIS is numerically unstable, so we adapted our results to adaptive optimisers. A few assumptions are needed to give meaningful results about their convergence (Défossez et al. 2020).

Assumption 3.4.

Assume that the ℓ_∞ norm of the gradient estimators of ρ and R, g and G respectively, are almost surely-bounded; that is $\exists R_1,\ R_2 \geq \sqrt{\varepsilon}$, such that $\forall x \in \mathbb{R}^d$:

$$\|g(x)\|_\infty \leq R_1 - \sqrt{\varepsilon} \quad \text{a.s.},$$
$$\|G(x)\|_\infty \leq R_2 - \sqrt{\varepsilon} \quad \text{a.s.}$$

Assumption 3.5.

Assume $\rho(\theta)$ is μ-strongly convex and L-smooth w.r.t. the $\|\cdot\|_\Theta$ norm.

Adam OAIS

If Assumption 3.4 holds on g and Assumption 3.5 holds on $\rho(\theta)$, using a constant learning rate $t_k = \alpha$ with parameters $\beta_2 \in (0, 1)$ and $\beta_1 \in (0, \beta_2)$ yields, for $T \geq \frac{\beta_1}{1-\beta_1}$:

Adam OAIS

If Assumption 3.4 holds on g and Assumption 3.5 holds on $\rho(\theta)$, using a constant learning rate $t_k = \alpha$ with parameters $\beta_2 \in (0,\, 1)$ and $\beta_1 \in (0,\, \beta_2)$ yields, for $T \geq \frac{\beta_1}{1-\beta_1}$:

Theorem 3.4.

$$\min_{k\in[T]_0} \mathbb{E}\left[|(\phi,\, \pi) - (\phi,\, \tilde{\pi}_k^N)|^2\right] \leq \frac{4R_1\|\phi\|_\infty^2}{N\alpha\mu}\frac{\rho(\theta_0)-\rho(\theta^*)}{\tilde{T}+1} + \frac{4E'\|\phi\|_\infty^2}{N(\tilde{T}+1)}\left[\log\left(1+\frac{R_1^2}{(1-\beta_2)\varepsilon}\right) - (T+1)\log(\beta_2)\right] + \frac{4\|\phi\|_\infty^2\rho(\theta^*)}{N}$$

Where $E' = E'(R_1,\, L,\, d,\, \alpha,\, \beta_1,\, \beta_2,\, \mu)$ and $\tilde{T} \propto T$ linearly.

Adam OAIS

If Assumption 3.4 holds on g and Assumption 3.5 holds on $\rho(\theta)$, using a constant learning rate $t_k = \alpha$ with parameters $\beta_2 \in (0,\, 1)$ and $\beta_1 \in (0,\, \beta_2)$ yields, for $T \geq \frac{\beta_1}{1-\beta_1}$:

Theorem 3.4.

$$\min_{k\in[T]_0} \mathbb{E}\left[|(\phi,\, \pi) - (\phi,\, \tilde{\pi}_k^N)|^2\right] \leq \frac{4R_1\|\phi\|_\infty^2}{N\alpha\mu}\frac{\rho(\theta_0)-\rho(\theta^*)}{\tilde{T}+1} + \frac{4E'\|\phi\|_\infty^2}{N(\tilde{T}+1)}\left[\log\left(1+\frac{R_1^2}{(1-\beta_2)\varepsilon}\right) - (T+1)\log(\beta_2)\right] + \frac{4\|\phi\|_\infty^2\rho(\theta^*)}{N}$$

Where $E' = E'(R_1,\, L,\, d,\, \alpha,\, \beta_1,\, \beta_2,\, \mu)$ and $\tilde{T} \propto T$ linearly.

Adam OAIS

If Assumption 3.4 holds on g and Assumption 3.5 holds on $\rho(\theta)$, using a constant learning rate $t_k = \alpha$ with parameters $\beta_2 \in (0,\, 1)$ and $\beta_1 \in (0,\, \beta_2)$ yields, for $T \geq \frac{\beta_1}{1-\beta_1}$:

Theorem 3.4.

$$\min_{k \in [T]_0} \mathbb{E}\left[|(\phi,\, \pi) - (\phi,\, \tilde{\pi}_k^N)|^2\right] \leq \frac{4R_1\|\phi\|_\infty^2}{N\alpha\mu}\frac{\rho(\theta_0) - \rho(\theta^*)}{\tilde{T}+1} + \frac{4E'\|\phi\|_\infty^2}{N(\tilde{T}+1)}\left[\log\left(1 + \frac{R_1^2}{(1-\beta_2)\varepsilon}\right) - (T+1)\log(\beta_2)\right] + \frac{4\|\phi\|_\infty^2\rho(\theta^*)}{N}$$

Where $E' = E'(R_1,\, L,\, d,\, \alpha,\, \beta_1,\, \beta_2,\, \mu)$ and $\tilde{T} \propto T$ linearly.

Remark

The bound is on the *minimum* MSE after T iterations

AdaGrad OAIS

If Assumption 3.4 holds on g and Assumption 3.5 holds on $\rho(\theta)$, then after T iterations with learning rate $t_k = \alpha$ and $\beta_1 \in (0, 1)$:

AdaGrad OAIS

If Assumption 3.4 holds on g and Assumption 3.5 holds on $\rho(\theta)$, then after T iterations with learning rate $t_k = \alpha$ and $\beta_1 \in (0, 1)$:

Theorem 3.6.

$$\min_{k \in [T]_0} \mathbb{E}\left[|(\phi, \pi) - (\phi, \tilde{\pi}_k^N)|^2 \right] \leq \frac{4R_1 \|\phi\|_\infty^2 [\rho(\theta_0) - \rho(\theta^*)]}{\mu \alpha N \sqrt{T+1}} + \frac{(8dR_1^2 + 2\alpha d R_1 L)\|\phi\|_\infty^2}{\mu N \sqrt{T+1}} \log\left(1 + \frac{(T+1)R_1^2}{\varepsilon}\right) + \frac{4\|\phi\|_\infty^2 \rho(\theta^*)}{N}$$

AdaGrad OAIS

If Assumption 3.4 holds on g and Assumption 3.5 holds on $\rho(\theta)$, then after T iterations with learning rate $t_k = \alpha$ and $\beta_1 \in (0, 1)$:

Theorem 3.6.

$$\min_{k \in [T]_0} \mathbb{E}\left[|(\phi, \pi) - (\phi, \tilde{\pi}_k^N)|^2 \right] \leq \frac{4R_1 \|\phi\|_\infty^2 [\rho(\theta_0) - \rho(\theta^*)]}{\mu \alpha N \sqrt{T+1}} + \frac{(8dR_1^2 + 2\alpha d R_1 L)\|\phi\|_\infty^2}{\mu N \sqrt{T+1}} \log\left(1 + \frac{(T+1)R_1^2}{\varepsilon}\right) + \frac{4\|\phi\|_\infty^2 \rho(\theta^*)}{N}$$

AdaGrad OAIS

Theorems 3.4 & 3.6 also are novel in the IS/OAIS setting.

AdaGrad OAIS

Theorems 3.4 & 3.6 also are novel in the IS/OAIS setting.

These were shown using convergence results for adaptive optimisers shown in (Défossez et al. 2020).

Contents

Preliminaries

Three main settings were considered:

Preliminaries

Three main settings were considered:

Case 1: π (Bivariate) Gaussian, q Gaussian – (Gaussian target)

Preliminaries

Three main settings were considered:

Case 1: π (Bivariate) Gaussian, q Gaussian – (Gaussian target)

Case 2: π Mixture Gaussian, q Gaussian – (Mixture target)

Preliminaries

Three main settings were considered:

Case 1: π (Bivariate) Gaussian, q Gaussian – (Gaussian target)

Case 2: π Mixture Gaussian, q Gaussian – (Mixture target)

Case 3: π Logit Normal, q Beta – (Logit Normal target)

Preliminaries

Three main settings were considered:

Case 1: π (Bivariate) Gaussian, q Gaussian – (Gaussian target)

Case 2: π Mixture Gaussian, q Gaussian – (Mixture target)

Case 3: π Logit Normal, q Beta – (Logit Normal target)

In all cases, $N = 1000$ atoms were used at each iteration to construct the empirical measure. 10 runs were performed in the first two cases, whilst 100 runs were performed in the last setting.

Preliminaries

Three main settings were considered:

Case 1: π (Bivariate) Gaussian, q Gaussian – (Gaussian target)

Case 2: π Mixture Gaussian, q Gaussian – (Mixture target)

Case 3: π Logit Normal, q Beta – (Logit Normal target)

In all cases, $N = 1000$ atoms were used at each iteration to construct the empirical measure. 10 runs were performed in the first two cases, whilst 100 runs were performed in the last setting.

Only *two* variables: number of iterations T and learning rate t_k (α if fixed).

Preliminaries

Three main settings were considered:

Case 1: π (Bivariate) Gaussian, q Gaussian – (Gaussian target)

Case 2: π Mixture Gaussian, q Gaussian – (Mixture target)

Case 3: π Logit Normal, q Beta – (Logit Normal target)

In all cases, $N = 1000$ atoms were used at each iteration to construct the empirical measure. 10 runs were performed in the first two cases, whilst 100 runs were performed in the last setting.

Only *two* variables: number of iterations T and learning rate t_k (α if fixed).

We will estimate $\mathbb{P}(X \in D)$ where $X \sim \pi$ and D will be specified. Equivalent to computing $(1_D, \pi)$.

Contents

Update rule

Let $q_{\theta_k} \sim \mathcal{N}(\mu_k, \Sigma_k)$. Defining $m_k = \Sigma_k^{-1}\mu_k$ and $S_k = \Sigma_k^{-1}$:

Update rule

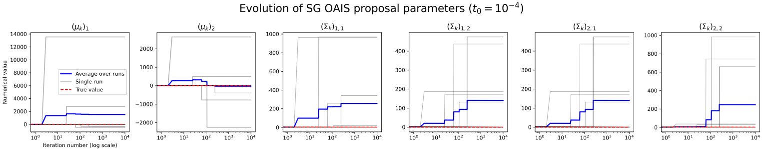
Let $q_{\theta_k} \sim \mathcal{N}(\mu_k, \Sigma_k)$. Defining $m_k = \Sigma_k^{-1}\mu_k$ and $S_k = \Sigma_k^{-1}$:

$$m_{k+1} \leftarrow m_k + \frac{t_k}{N} \sum_{i=1}^{N} \frac{\pi^2(x_i)}{q^2_{\theta_k}(x_i)}(x_i - S_k^{-1} m_k)$$

$$S_{k+1} \leftarrow \mathrm{Proj}_{\mathrm{PD}^2}\left[S_k - \frac{t_k}{2N} \sum_{i=1}^{N} \frac{\pi^2(x_i)}{q^2_{\theta_k}(x_i)}(x_i x_i^\top - S_k^{-1} m_k m_k^\top S_k^{-1} - S_k^{-1})\right]$$

SG OAIS (Gaussian Target)

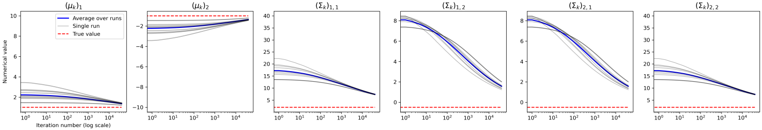
$$T = 10000,\ t_k = \frac{10^{-4}}{\sqrt{k+1}}$$



SG OAIS (Gaussian Target)

$$T = {\color{red}40000},\ t_k = \frac{10^{-{\color{red}5}}}{\sqrt{k+1}}$$

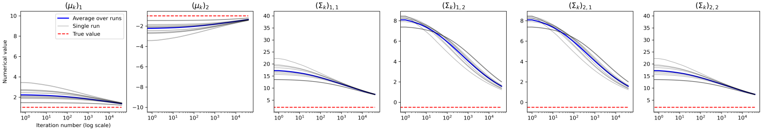
Evolution of SG OAIS proposal parameters



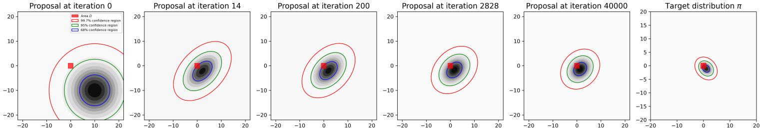
SG OAIS (Gaussian Target)

$$T = 40000,\ t_k = \frac{10^{-5}}{\sqrt{k+1}}$$

Evolution of SG OAIS proposal parameters



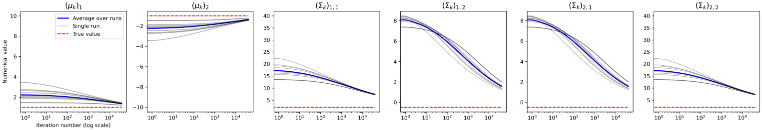
Evolution of SGD OAIS distribution (average over 10 experiments)



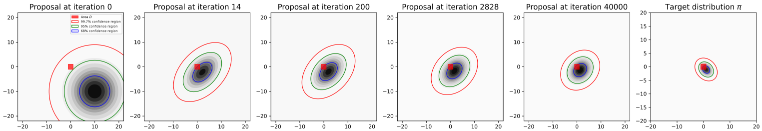
SG OAIS (Gaussian Target)

$$T = 40000, \; t_k = \frac{10^{-5}}{\sqrt{k+1}}$$

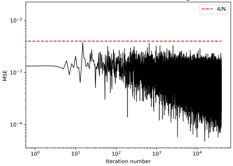
Evolution of SG OAIS proposal parameters



Evolution of SGD OAIS distribution (average over 10 experiments)

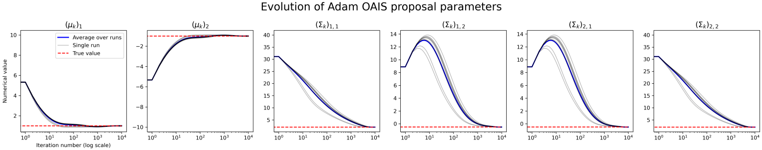


MSE of SG OAIS (10 run average)



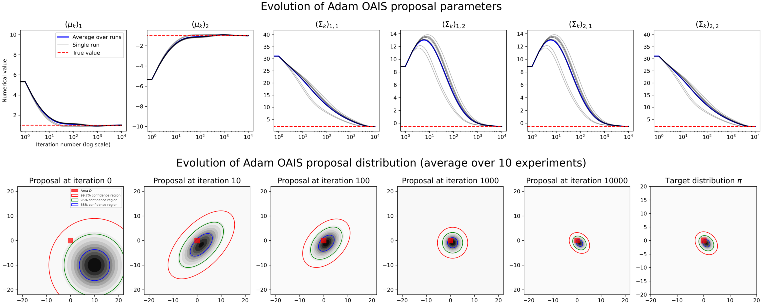
Adam OAIS (Gaussian Target)

$$T = 10000,\ t_k = \alpha = 0.01$$



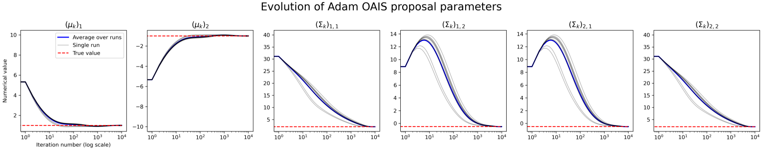
Adam OAIS (Gaussian Target)

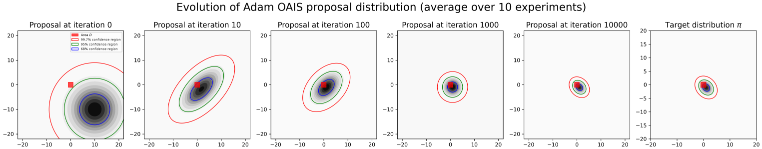
$$T = 10000,\ t_k = \alpha = 0.01$$

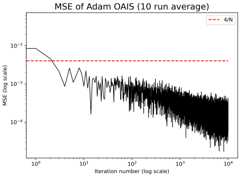


Adam OAIS (Gaussian Target)

$$T = 10000, \; t_k = \alpha = 0.01$$

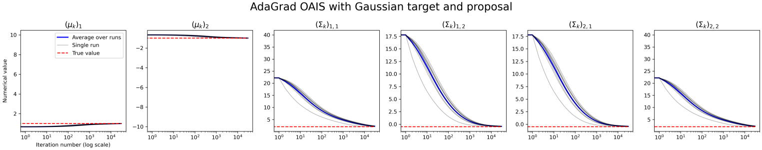


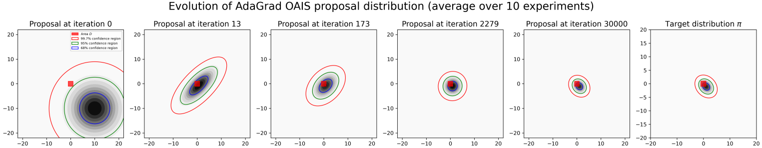


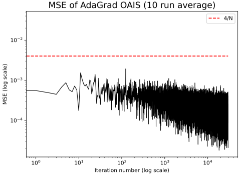


AdaGrad OAIS (Gaussian Target)

$$T = 30000,\ t_k = \alpha = 0.1$$





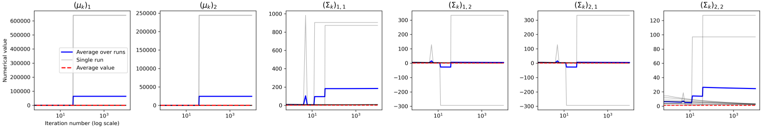


Contents

SG OAIS (Mixture Target)

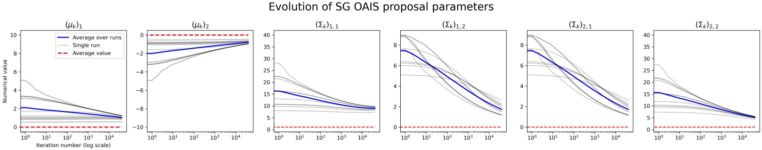
$$T = 10000,\ t_k = \frac{10^{-4}}{\sqrt{k+1}}$$

Evolution of SG OAIS proposal parameters ($t_0 = 10^{-4}$)



SG OAIS (Mixture Target)

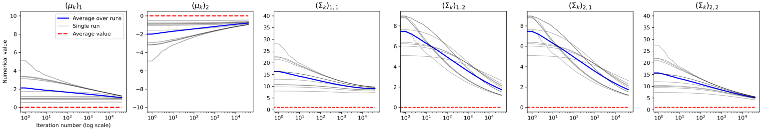
$$T = 40000, \ t_k = \frac{10^{-5}}{\sqrt{k+1}}$$



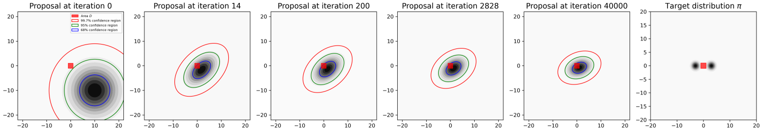
SG OAIS (Mixture Target)

$$T = 40000,\ t_k = \frac{10^{-5}}{\sqrt{k+1}}$$

Evolution of SG OAIS proposal parameters



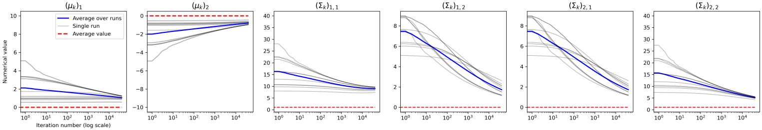
Evolution of SG OAIS proposal distribution (average over 10 experiments)



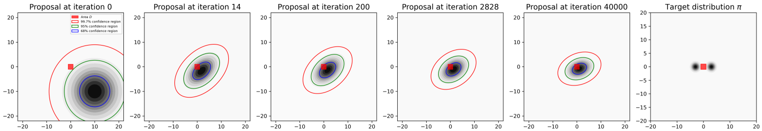
SG OAIS (Mixture Target)

$$T = 40000,\ t_k = \frac{10^{-5}}{\sqrt{k+1}}$$

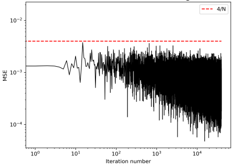
Evolution of SG OAIS proposal parameters



Evolution of SG OAIS proposal distribution (average over 10 experiments)

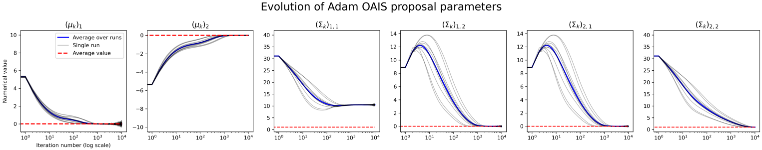


MSE of SG OAIS (10 run average)



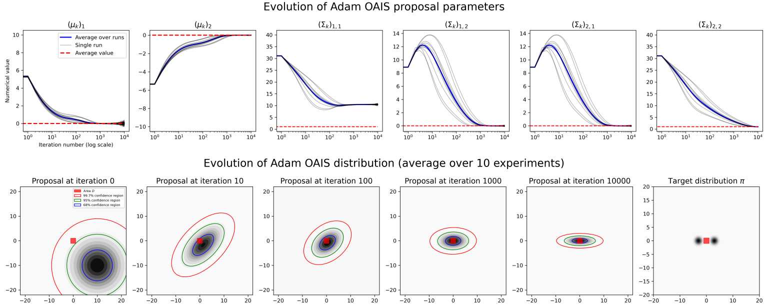
Adam OAIS (Mixture Target)

$$T = 10000,\ t_k = \alpha = 0.01$$



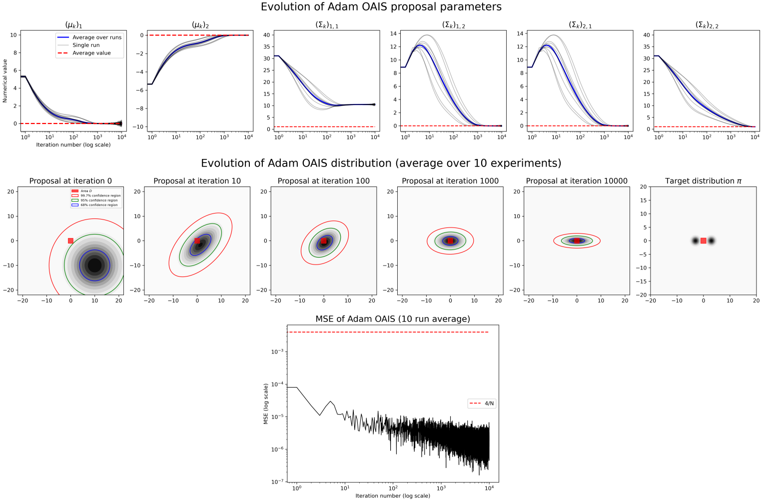
Adam OAIS (Mixture Target)

$$T = 10000, \; t_k = \alpha = 0.01$$



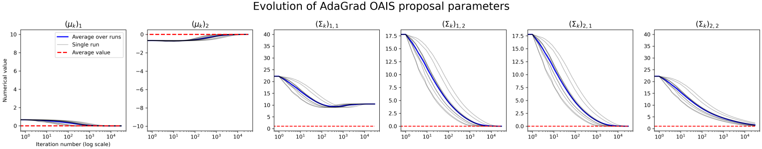
Adam OAIS (Mixture Target)

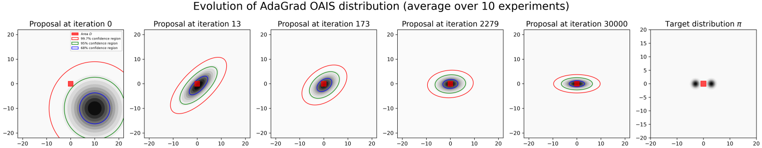
$$T = 10000, \; t_k = \alpha = 0.01$$

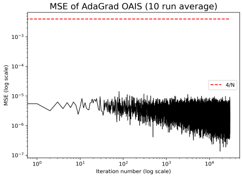


AdaGrad OAIS (Mixture Target)

$$T = 30000, \ t_k = \alpha = 0.1$$







Contents

Logit Normal

If $X \sim \mathcal{N}(\mu, \sigma)$, $\frac{\exp(X)}{1+\exp(X)} := Y \sim \text{LogitNormal}(\mu, \sigma)$.

Logit Normal

If $X \sim \mathcal{N}(\mu, \sigma)$, $\frac{\exp(X)}{1+\exp(X)} := Y \sim \text{LogitNormal}(\mu, \sigma)$. Highly intractable: $E[Y] = 0.5$ if $\mu = 0$, all other moments and cases unknown (Holmes and Schofield 2022).

Logit Normal

If $X \sim \mathcal{N}(\mu, \sigma)$, $\frac{\exp(X)}{1+\exp(X)} := Y \sim \text{LogitNormal}(\mu, \sigma)$. Highly intractable: $E[Y] = 0.5$ if $\mu = 0$, all other moments and cases unknown (Holmes and Schofield 2022).

What proposal q_θ to use?

Logit Normal

If $X \sim \mathcal{N}(\mu, \sigma)$, $\frac{\exp(X)}{1+\exp(X)} := Y \sim \text{LogitNormal}(\mu, \sigma)$. Highly intractable: $E[Y] = 0.5$ if $\mu = 0$, all other moments and cases unknown (Holmes and Schofield 2022).

What proposal q_θ to use? $\text{Supp}(Y) = (0, 1)$

Logit Normal

If $X \sim \mathcal{N}(\mu, \sigma)$, $\frac{\exp(X)}{1+\exp(X)} := Y \sim \text{LogitNormal}(\mu, \sigma)$. Highly intractable: $E[Y] = 0.5$ if $\mu = 0$, all other moments and cases unknown (Holmes and Schofield 2022).

What proposal q_θ to use? $\text{Supp}(Y) = (0, 1) \rightsquigarrow$ Beta proposal

Logit Normal

If $X \sim \mathcal{N}(\mu, \sigma)$, $\frac{\exp(X)}{1+\exp(X)} := Y \sim \text{LogitNormal}(\mu, \sigma)$. Highly intractable: $E[Y] = 0.5$ if $\mu = 0$, all other moments and cases unknown (Holmes and Schofield 2022).

What proposal q_θ to use? $\text{Supp}(Y) = (0, 1) \rightsquigarrow$ Beta proposal

$$\alpha_{k+1} \leftarrow \left| \alpha_k + \frac{t_k}{N} \sum_{i=1}^{N} \frac{\pi^2(x_i)}{q^2_{\theta_k}(x_i)} \left[\psi^0(\alpha_k + \beta_k) - \psi^0(\alpha_k) + \log(x_i) \right] \right|$$

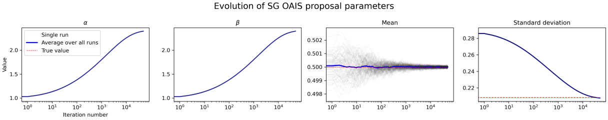
$$\beta_{k+1} \leftarrow \left| \beta_k + \frac{t_k}{N} \sum_{i=1}^{N} \frac{\pi^2(x_i)}{q^2_{\theta_k}(x_i)} \left[\psi^0(\alpha_k + \beta_k) - \psi^0(\beta_k) + \log(1 - x_i) \right] \right|$$

Where x_i i.i.d. and $x_i \sim q_{\theta_k}$ and $\psi^0(x)$ is the Digamma function.

SG OAIS (Logit Normal Target)

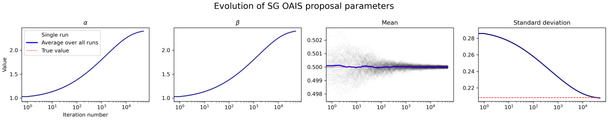
SG OAIS (Logit Normal Target)

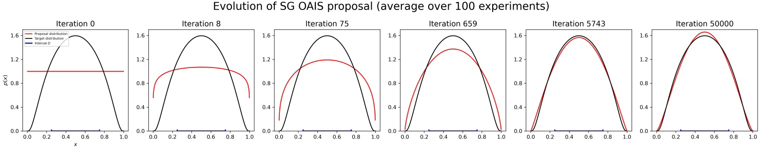
$$T = 50000,\ t_k = \frac{10}{\sqrt{k+1}}$$



SG OAIS (Logit Normal Target)

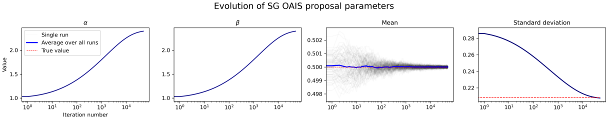
$$T = 50000, \; t_k = \frac{10}{\sqrt{k+1}}$$

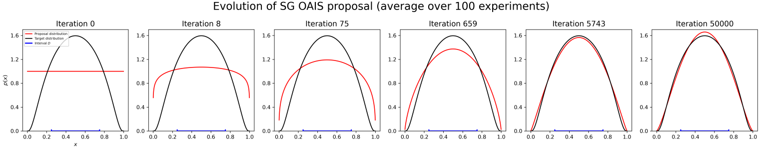


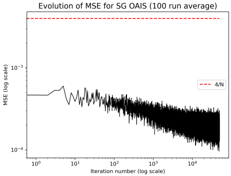


SG OAIS (Logit Normal Target)

$$T = 50000,\ t_k = \frac{10}{\sqrt{k+1}}$$

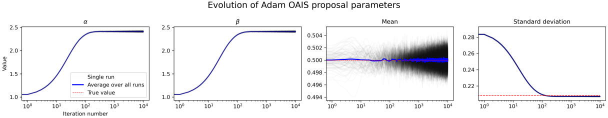






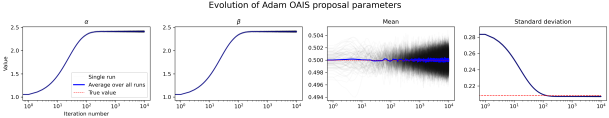
Adam OAIS (Logit Normal Target)

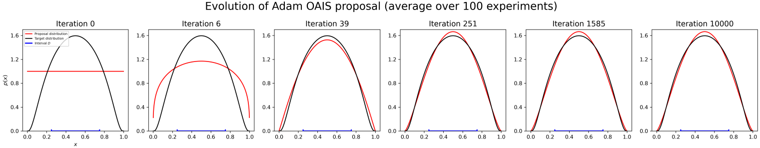
$$T = 10000, \; t_k = \alpha = 0.1$$



Adam OAIS (Logit Normal Target)

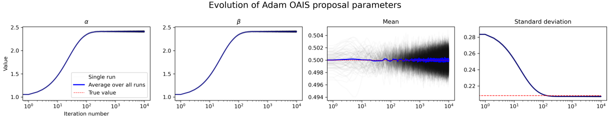
$$T = 10000, \; t_k = \alpha = 0.1$$

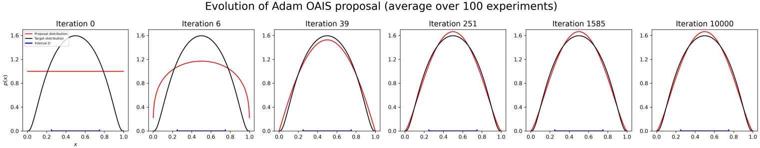


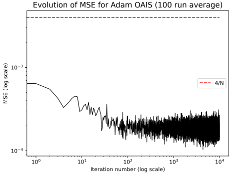


Adam OAIS (Logit Normal Target)

$$T = 10000, \; t_k = \alpha = 0.1$$



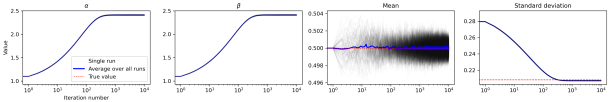




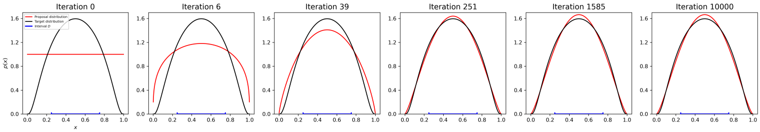
AdaGrad OAIS (Logit Normal Target)

$$T = 10000, \; t_k = \alpha = 0.01$$

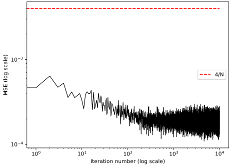
Evolution of AdaGrad OAIS proposal parameters



Evolution of AdaGrad OAIS proposal (average over 100 experiments)



Evolution of MSE for AdaGrad OAIS (100 run average)



Summary

Method	Assumptions	Convexity	Type of Bound	Adaptive Rate	Reference
SG OAIS	2.1, 2.2, 3.1, 3.2	Regular	Last iterate	$\mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right)$	Theorem 3.2
USG OAIS	2.1, 2.2, 3.1, 3.2	Regular	Last iterate	$\mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right)$	Theorem 3.2
Adam OAIS	3.4, 3.5	Strong	Min-iterate	—	Theorem 3.4
AdaGrad OAIS	3.4, 3.5	Strong	Min-iterate	$\mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right)$	Theorem 3.6

Table 1: The OAIS algorithms and their convergence rates in unbounded parameter domains.

Bibliography

Akyildiz, Ö. D. and J. Míguez (2021). "Convergence rates for optimised adaptive importance samplers". In: *Statistics and Computing* 31.2. ISSN: 15731375. DOI: 10.1007/s11222-020-09983-1.

Orabona, F. (Aug. 2020). *Last Iterate of SGD Converges (Even in Unbounded Domains)*. URL: https://parameterfree.com/2020/08/07/last-iterate-of-sgd-converges-even-in-unbounded-domains/.

Défossez, A. et al. (2020). "A Simple Convergence Proof of Adam and Adagrad". In: *Trans. Mach. Learn. Res.* 2022.

Holmes, J. B. and M. R. Schofield (Feb. 2022). "Moments of the logit-normal distribution". In: *Communications in Statistics - Theory and Methods* 51.3, pp. 610–623. ISSN: 0361-0926. DOI: 10.1080/03610926.2020.1752723. URL: https://doi.org/10.1080/03610926.2020.1752723.