Adaptive Multilevel Monte Carlo for Probabilities

Citation for published version:
Haji-Ali, A-L, Spence, J & Teckentrup, A 2022, 'Adaptive Multilevel Monte Carlo for Probabilities', *SIAM Journal on Numerical Analysis*, vol. 60, no. 4, pp. 2125-2149. https://doi.org/10.1137/21M1447064

Digital Object Identifier (DOI):
10.1137/21M1447064

Published In:
SIAM Journal on Numerical Analysis

ADAPTIVE MULTILEVEL MONTE CARLO FOR PROBABILITIES*

ABDUL-LATEEF HAJI-ALI†, JONATHAN SPENCE†, AND ARETHA TECKENTRUP‡

Abstract. We consider the numerical approximation of $\mathbb{P}[\, G \in \Omega \,]$ where the d-dimensional random variable G cannot be sampled directly, but there is a hierarchy of increasingly accurate approximations $\{G_\ell\}_{\ell\in\mathbb{N}}$ which can be sampled. The cost of standard Monte Carlo estimation scales poorly with accuracy in this setup since it compounds the approximation and sampling cost. A direct application of Multilevel Monte Carlo improves this cost scaling slightly, but returns sub-optimal computational complexities since estimation of the probability involves a discontinuous functional of G_ℓ. We propose a general adaptive framework which is able to return the MLMC complexities seen for smooth or Lipschitz functionals of G_ℓ. Our assumptions and numerical analysis are kept general allowing the methods to be used for a wide class of problems. We present numerical experiments on nested simulation for risk estimation, where $G = \mathbb{E}[\, X|Y \,]$ is approximated by an inner Monte Carlo estimate. Further experiments are given for digital option pricing, involving an approximation of a d-dimensional SDE.

AMS subject classifications. 65C05, 62P05

Key words. Multilevel Monte Carlo, Nested simulation, Risk estimation

1. Introduction.

This paper proposes general, efficient numerical methods to compute

$$\mathbb{P}[\, G \in \Omega \,] = \mathbb{E}[\, \mathbb{I}_{G\in\Omega} \,], \qquad \mathbb{I}_{G\in\Omega} \coloneqq \begin{cases} 1 & G \in \Omega \\ 0 & G \notin \Omega \end{cases}, \tag{1.1}$$

within an error tolerance ε, where G is a d-dimensional random variable which cannot be sampled directly and $\mathbb{I}_{G\in\Omega}$ is the indicator of the set Ω. In Subsection 1.1, we relate (1.1) to the one-dimensional problem

$$\mathbb{P}[\, g > 0 \,] = \mathbb{E}[\, \mathbb{H}(g) \,], \tag{1.2}$$

where $\mathbb{H}(g)$ is the Heaviside function, equal to 1 when $g \geq 0$ and to 0 otherwise. In most problems of interest, g requires approximate sampling. We assume access to a hierarchy of increasingly accurate approximations $\{g_\ell\}_{\ell\in\mathbb{N}}$ converging to g almost surely as $\ell \to \infty$. Approximate simulation of g induces a bias in typical Monte Carlo methods for (1.2), increasing the cost of standard Monte Carlo averages. In such situations, Multilevel Monte Carlo (MLMC) [6, 11, 12] is often able to reduce the cost, but is known to suffer when the observable is discontinuous as in (1.1) or (1.2) [9, 10, 13]. Adaptive sampling techniques [5, 9, 13] have proven successful in reducing the cost of Monte Carlo and MLMC for specific instances of (1.2). This paper builds upon such methods to establish a general framework for this problem with an emphasis on ensuring applicability to wide ranging problems. Examples are discussed below.

EXAMPLE 1.1 (Nested Simulation). *Equation* (1.2) *often arises in financial risk estimation. For example, many risk measures involve conditional expectations of the form $g = \mathbb{E}[\, X|Y \,]$ for some random variables X, Y [13, 14, 18, 19]. Approximation of g by g_ℓ is possible using an inner Monte Carlo average with $N_\ell \in \mathbb{N}$ samples.*

*Submitted 17 September, 2021

†Heriot-Watt University (a.hajiali@hw.ac.uk, jws5@hw.ac.uk).
‡University of Edinburgh (a.teckentrup@ed.ac.uk)

EXAMPLE 1.2 (Digital Option Pricing). *Let S satisfy the d-dimensional SDE*

$$dS(t) = a(t, S(t))\,dt + b(t, S(t))\,dW(t),$$

for sufficiently smooth functions a and b, and Brownian motion W. For a (European-type) digital option with deterministic maturity $T > 0$, we set $G \equiv S(T)$ and return a unit payoff if $G \in \Omega$ and no payoff otherwise. The (non-discounted) value at time 0 of this option is of the form (1.1)*, where G can be approximately sampled using SDE discretisation methods [21].*

A related setup is discussed in [9] and applied in [8] to compute failure properties of systems governed by PDEs. In [9], the idea of selective refinement is used to adaptively refine MLMC samples based on the uncertainty of $g > 0$. Selective refinement aims to reduce the cost of sampling level ℓ without affecting the approximation error of $\mathbb{H}(g)$. There, it is assumed that the error $|g - g_\ell|$ is bounded when g_ℓ is near zero, excluding applications like Examples 1.1 and 1.2.

There is extensive research into Monte Carlo approximation of nested simulation problems as in Example 1.1. Analysis of standard Monte Carlo methods for nested simulation is discussed in [18]. Adaptivity is then combined with standard Monte Carlo methods for this problem in [5]. Moreover, in [13, 14] adaptive MLMC methods for nested simulation are discussed. Contrary to the selective refinement algorithm in [9], these methods aim to improve the approximation error of $\mathbb{H}(g)$ at level ℓ while the average work of sampling at level ℓ is relatively unaffected. This approach forms the basis for the present work. Similar results are obtained in [24], where the authors approximate the inner expectation $\mathbb{E}[\,X|Y\,]$ using Quasi-Monte Carlo techniques.

An alternative approach to compute (1.2) via MLMC is to approximate $\mathbb{H}(g)$ by a Lipschitz function. This idea of smoothing the Heaviside function has been employed successfully in [2, 15, 23]. See also [3], where a polynomial chaos expansion is used to approximate the indicator function of a random variable. These approaches require an explicit smoothing step, which the work presented here removes by using adaptivity.

The key contributions of this paper are as follows:

- A generalisation of the adaptive MLMC sampling scheme for nested simulation [13, 14] is presented in Algorithm 3.1. The new procedure requires less restrictive moment bounds on g and is formulated in a general framework allowing for applications beyond nested simulation.
- By reformulating the ideas in [13], we are able to significantly simplify the analysis compared with the previous work.
- Numerical experiments show the adaptive MLMC scheme introduced here remains effective for nested simulation, with a slight relaxation of the sampling process used in [13, 14]. Additional results show the scheme has an equally strong impact when applied to digital option pricing as in Example 1.2.

Subsection 1.1 outlines the problem setup and necessary assumptions for this analysis, before discussing the link between problems (1.1) and (1.2). We describe the MLMC approach to (1.2) in Section 2 and show how the complexity of MLMC suffers because $\mathbb{H}(g)$ is discontinuous. We show how the results can be improved slightly under stronger assumptions. In Section 3, we introduce the adaptive MLMC procedure and analyse its benefits to the MLMC complexity. Numerical results are then presented in Section 4.

1.1. Problem Setup. For the majority of this paper, we focus on the problem (1.2). At the end of this section, we discuss how to extend the methods to general problems of the form (1.1). As is typical for MLMC, we assume the expected sampling cost of g_ℓ, denoted W_ℓ, increases geometrically with ℓ. For ease of notation we use the operator $f_0 \lesssim f_1$ throughout this paper to denote $f_0 \le C \cdot f_1$, where C is independent of ℓ and the error tolerance ε. In particular,

$$W_\ell \lesssim 2^{\gamma \ell}, \quad \text{for some } \gamma > 0. \tag{1.3}$$

The following assumption controls the strong approximation error of g_ℓ.

Assumption 1.3. *For some $2 < q$, $\beta > 0$ and positive valued random variable σ_ℓ, define*

$$Z_\ell \coloneqq \frac{g_\ell - g}{\sigma_\ell 2^{-\beta\ell/2}}, \tag{1.4}$$

and assume $\mathbb{E}[\,|Z_\ell|^q\,]$ is uniformly bounded in $\ell \ge 0$.

In this context, σ_ℓ represents fluctuations in the approximation uncertainty for a given instance of g_ℓ. In practice, σ_ℓ will typically form an estimate of the variability of g_ℓ. For example, in the nested simulation problem (Example 1.1), where $g = \mathbb{E}[\,X|Y\,]$, we can take σ_ℓ to be the sample standard deviation of X given Y, as in [13]. Assumption 1.3 allows us to use Markov's inequality to bound

$$\mathbb{P}[\,|Z_\ell| \ge x\,] \le x^{-q}\mathbb{E}[\,|Z_\ell|^q\,] \tag{1.5}$$

for all $x > 0$. This result is used in many proofs within this paper.

To implement MLMC successfully, we control the probability of sampling g_ℓ close to 0. In doing so, we introduce the parameter

$$\delta_\ell \coloneqq \frac{g_\ell}{\sigma_\ell}, \tag{1.6}$$

which models the sample specific uncertainty in the sign of g_ℓ and thus $\mathbb{H}(g_\ell)$.

Assumption 1.4. *There exists $\delta, \rho_0 > 0$ such that for all $0 < x \le \delta$ we have*

$$\mathbb{P}[\,|\delta_\ell| < x\,] \le \rho_0 x$$

for all $\ell \ge 0$.

Assumptions 1.3 and 1.4 are enough to bound the strong error of approximations $\mathbb{H}(g_\ell)$, which underpin the complexity theory for MLMC approximation of (1.2).

It is important to remark here that the assumptions above allow for the simple extension to the general problem (1.1) under equivalent assumptions. To see this, assume that (for $\|\cdot\|$ being the Euclidean norm)

$$d_\Omega(G) \coloneqq \min_{\omega \in \partial\Omega} \{\|G - \omega\|\}$$

exists. Here, we are assuming the minimum distance to the boundary of Ω is attained by a point on the boundary. Then, (1.1) is equivalent to (1.2) when

$$g = \bar{d}_\Omega(G) \coloneqq \begin{cases} d_\Omega(G) & G \in \Omega \\ -d_\Omega(G) & G \notin \Omega \end{cases}$$

is a signed distance. If we denote approximations of G at level $\ell \in \mathbb{N}$ by G_ℓ then we have approximations $g_\ell \coloneqq \bar{d}_\Omega(G_\ell)$ of g. Assumption 1.3 then holds provided

$$\mathbb{E}\left[\left(\frac{\|G - G_\ell\|}{\sigma_\ell 2^{-\beta\ell/2}}\right)^q\right]$$

is uniformly bounded in ℓ, since the Euclidean norm is Lipschitz continuous. Assumption 1.4 becomes an equivalent condition on the distribution of $|\delta_\ell| = d_\Omega(G_\ell)/\sigma_\ell$.

2. Multilevel Monte Carlo for Probabilities. In this section, we outline the use of standard MLMC methods [6, 11, 12] for approximating (1.2). In particular, we show that the discontinuity at 0 in the Heaviside function limits the effectiveness of standard MLMC for this problem. Similar arguments from the context of nested simulation can be found in [13, 16, 18]. We begin by approximating $\mathbb{P}[\,g > 0\,]$ by $\mathbb{P}[\,g_L > 0\,]$, where L should be chosen large enough to control the approximation bias. Sampling g at large levels L is typically expensive. The key idea of MLMC is to split this computation over levels $0 \le \ell \le L$ using a telescopic sum. Specifically, using $\mathbb{H}(g_{-1}) \coloneqq 0$

$$\begin{aligned} \mathbb{E}[\,\mathbb{H}(g)\,] \approx \mathbb{E}[\,\mathbb{H}(g_L)\,] &= \sum_{\ell=0}^{L} \mathbb{E}[\,\mathbb{H}(g_\ell) - \mathbb{H}(g_{\ell-1})\,] \\ &\approx \sum_{\ell=0}^{L} \left(\frac{1}{M_\ell} \sum_{m=1}^{M_\ell} \Big(\mathbb{H}(g_\ell^{(f,m)}) - \mathbb{H}(g_{\ell-1}^{(c,m)}) \Big) \right), \end{aligned} \tag{2.1}$$

where we approximate each expectation in the telescopic sum by an independent Monte Carlo sum with samples $\mathbb{H}(g_\ell^{(f,m)}) - \mathbb{H}(g_{\ell-1}^{(c,m)}) \overset{\text{i.i.d}}{\sim} \mathbb{H}(g_\ell) - \mathbb{H}(g_{\ell-1})$. Samples $g_\ell^{(f,m)}$ and $g_{\ell-1}^{(c,m)}$ should be closely correlated to reduce $\mathrm{Var}[\,\mathbb{H}(g_\ell) - \mathbb{H}(g_{\ell-1})\,]$, lowering the number of samples, M_ℓ required at level ℓ. The following result bounds the total work of sampling (2.1) within a given error tolerance.

Proposition 2.1 ([6, 12]). *Let $\{\Delta\mathbb{H}_\ell\}_{\ell=0}^{\infty}$ be a sequence of random variables with $\mathbb{P}[\,g > 0\,] = \sum_{\ell=0}^{\infty} \mathbb{E}[\,\Delta\mathbb{H}_\ell\,]$. Assume the following rates of convergence for some $\gamma, \beta_{\mathrm{ind}} > 0,\ \alpha_{\mathrm{ind}} \ge \frac{\min(\gamma, \beta_{\mathrm{ind}})}{2}$:*
- *The expected work of sampling $\Delta\mathbb{H}_\ell$ is $W_\ell \lesssim 2^{\gamma\ell}$.*
- *The mean and variance of $\Delta\mathbb{H}_\ell$ converge to 0 with the following rates*

$$E_\ell \coloneqq |\mathbb{E}[\,\Delta\mathbb{H}_\ell\,]| \lesssim 2^{-\alpha_{\mathrm{ind}}\ell}. \tag{2.2}$$

$$V_\ell \coloneqq \mathrm{Var}[\,\Delta\mathbb{H}_\ell\,] \lesssim 2^{-\beta_{\mathrm{ind}}\ell}, \tag{2.3}$$

Then, there is optimal L and $\{M_\ell\}_{0\le\ell\le L}$ such that the total work of computing the MLMC estimator

$$\mathcal{M}^L_{M_0,\dots,M_L} \coloneqq \sum_{\ell=0}^{L} \left(\frac{1}{M_\ell} \sum_{m=1}^{M_\ell} \Delta\mathbb{H}_\ell^{(m)} \right), \quad \Delta\mathbb{H}_\ell^{(m)} \overset{\text{i.i.d.}}{\sim} \Delta\mathbb{H}_\ell \tag{2.4}$$

with mean square error satisfying $\mathbb{E}[\,(\mathbb{P}[\,g > 0\,] - \mathcal{M}^L_{M_0,\dots,M_L})^2\,] \le \varepsilon^2$ is

$$\mathrm{Work}\big(\mathcal{M}^L_{M_0,\dots,M_L}, \varepsilon\big) \lesssim \begin{cases} \varepsilon^{-2} & \beta_{\mathrm{ind}} > \gamma \\ \varepsilon^{-2}(\log\varepsilon)^2 & \beta_{\mathrm{ind}} = \gamma \\ \varepsilon^{-2-(\gamma-\beta_{\mathrm{ind}})/\alpha_{\mathrm{ind}}} & \beta_{\mathrm{ind}} < \gamma \end{cases}.$$

We will denote the estimator (2.4) with optimal L and $\{M_\ell\}_{0\le\ell\le L}$ by $\mathcal{M}^*$.

Remark 2.2. *Proposition 2.1 can be applied to the MLMC estimator (2.1) by taking $\Delta\mathbb{H}_\ell \coloneqq \mathbb{H}(g_\ell) - \mathbb{H}(g_{\ell-1})$. In Section 3 we see $\Delta\mathbb{H}_\ell$ take a slightly different form to* *accommodate adaptive approximation of g. E_ℓ in (2.2) and V_ℓ in (2.3) are the bias and variance of the multilevel correction, respectively. Rather than prove convergence rates for these terms directly, we provide stronger results on $\mathbb{E}[\,(\mathbb{H}(g)-\mathbb{H}(g_\ell))^2\,], |\mathbb{E}[\,\mathbb{H}(g)-\mathbb{H}(g_\ell)\,]|$. The bound on $\mathrm{Work}(\mathcal{M}^*;\varepsilon)$ is sometimes referred to as the complexity of $\mathcal{M}^*$, since it describe how the total work scales as the error decreases. Replacing $\mathbb{H}(\cdot)$ with a smooth/Lipschitz functional, a similar result to Proposition 2.1 holds [6, 12] for $\beta_{\text{ind}}=\beta$ and we see ε^{-2} complexity for $\beta>\gamma$, up to an additional bias induced by the smoothing. In this paper, we refer to ε^{-2} as the 'canonical' complexity since it is the same as seen for standard Monte Carlo with exact sampling of g.*

The following result provides a bound on $\mathbb{E}[\,(\mathbb{H}(g_\ell)-\mathbb{H}(g_{\ell-1}))^2\,]$ under the assumptions in Subsection 1.1. The rate is worse than that of smooth/Lipschitz functionals mentioned in Remark 2.2, since we make an $\mathcal{O}(1)$ approximation error in $\mathbb{H}(g)-\mathbb{H}(g_\ell)$ whenever g, g_ℓ lie on opposite sides of 0.

PROPOSITION 2.3 (Variance With General Assumptions). *By Assumptions 1.3 and 1.4 we have $\mathbb{E}[\,(\mathbb{H}(g)-\mathbb{H}(g_\ell))^2\,]\lesssim 2^{-(\frac{q}{q+1})\ell\beta/2}$.*

Proof. We compute

$$\begin{aligned}\mathbb{E}[\,(\mathbb{H}(g)-\mathbb{H}(g_\ell))^2\,] &\le \mathbb{E}[\,\mathbb{I}_{|g-g_\ell|\ge|g_\ell|}\,]\\ &= \mathbb{E}[\,\mathbb{I}_{|Z_\ell|\ge 2^{\ell\beta/2}|\delta_\ell|}\,],\end{aligned}$$

where Z_ℓ and δ_ℓ are as in (1.4) and (1.6). It follows from Markov's inequality (1.5) that, for any $\psi>0$

$$\begin{aligned}\mathbb{E}[\,\mathbb{I}_{|Z_\ell|\ge 2^{\ell\beta/2}|\delta_\ell|}\,] &= \mathbb{E}[\,\mathbb{I}_{|Z_\ell|\ge 2^{\ell\beta/2}|\delta_\ell|}\mathbb{I}_{|\delta_\ell|\le\psi}\,]+\mathbb{E}[\,\mathbb{I}_{|Z_\ell|\ge 2^{\ell\beta/2}|\delta_\ell|}\mathbb{I}_{|\delta_\ell|\ge\psi}\,]\\ &\le \mathbb{E}[\,\mathbb{I}_{|\delta_\ell|\le\psi}\,]+\mathbb{E}[\,\mathbb{I}_{|Z_\ell|\ge 2^{\ell\beta/2}\psi}\,]\\ &\le \rho_0\psi+(2^{\ell\beta/2}\psi)^{-q}\mathbb{E}[\,|Z_\ell|^q\,],\end{aligned}$$

where we have used Assumption 1.4. Then we set $\psi=\min(1,\delta)2^{-(\frac{q}{q+1})\ell\beta/2}$ to get the previous two terms of equal rate, which is the variance convergence rate. ∎

REMARK 2.4. *Proposition 2.3 also proves an upper bound on E_ℓ for $\Delta\mathbb{H}_\ell=\mathbb{H}(g_\ell)-\mathbb{H}(g_{\ell-1})$ (2.2) since we have $|\mathbb{E}[\,\mathbb{H}(g_\ell)-\mathbb{H}(g_{\ell-1})\,]|\le\mathbb{E}[\,|\mathbb{H}(g_\ell)-\mathbb{H}(g_{\ell-1})|\,]\le \mathbb{E}\Big[\,(\mathbb{H}(g_\ell)-\mathbb{H}(g_{\ell-1}))^2\,\Big]$.*

REMARK 2.5. *All even moments of $\mathbb{H}(g_\ell)-\mathbb{H}(g_{\ell-1})$ are equal, thus Proposition 2.3 actually proves a bound for all absolute moments of $\mathbb{H}(g_\ell)-\mathbb{H}(g_{\ell-1})$. This leads to a large kurtosis of the multilevel correction which can impact the robustness of MLMC and is discussed further in Section 4.*

In the context of Proposition 2.1, Proposition 2.3 shows $\beta_{\text{ind}}=(\frac{q}{q+1})\frac{\beta}{2}$ and we only observe ε^{-2} complexity when $\beta>2(\frac{q+1}{q})\gamma$. In many examples, including those discussed here, $\beta\le 2\gamma$ and we need tight bounds on E_ℓ (2.2) to state accurate complexities. To derive tighter bounds than Remark 2.4 we require further assumptions.

ASSUMPTION 2.6 ([18]). *Let $\rho_\ell(y,z)$ be the joint density of δ_ℓ (1.6) and Z_ℓ (1.4), defined for some $\beta>0$. Assume that for all ℓ, ρ_ℓ is twice differentiable in y and there exists $p_{i,\ell}(\cdot)$ such that*

$$\left|\frac{\partial^i}{\partial y^i}\rho_\ell(y,z)\right|\le p_{i,\ell}(z),\quad \sup_\ell\int_{\mathbb{R}}|z|^j p_{i,\ell}(z)\,dz<\infty,$$

for $i=0,1,2$ and $0\le j\le q+2$ for some $q>2$.

ASSUMPTION 2.7. *For Z_ℓ, β as in Assumption 1.3, we have $|\mathbb{E}[\, Z_\ell \,]| \lesssim 2^{\ell(\beta/2-\alpha)}$, for some $\frac{\beta}{2} \le \alpha \le \beta$.*

From (1.4) we see that Assumption 2.7 bounds $\mathbb{E}\big[\, \sigma_\ell^{-1}(g-g_\ell) \,\big] \lesssim 2^{-\alpha\ell}$. Assumption 2.7 is instead expressed in terms of Z_ℓ to align with the analysis in Section 3 (see Assumption 3.8). We stress that these assumptions are required only to obtain better convergence rates of E_ℓ. Reasonable results can still be obtained using Remark 2.4 when they are false. Nonetheless, Assumption 2.6 also provides slightly better bounds for $\mathbb{E}[\, (\mathbb{H}(g_\ell) - \mathbb{H}(g_{\ell-1}))^2 \,]$. For completeness, we state this result below.

PROPOSITION 2.8 (Variance With Strict Assumptions). *Under Assumption 2.6 it follows that $\mathbb{E}\Big[\, (\mathbb{H}(g) - \mathbb{H}(g_\ell))^2 \,\Big] \lesssim 2^{-\ell\beta/2}$.*

Proof. By Assumption 2.6,we have

$$\begin{aligned}
\mathbb{E}\Big[\, (\mathbb{H}(g) - \mathbb{H}(g_\ell))^2 \,\Big] &\le \mathbb{E}\big[\, \mathbb{I}_{|Z_\ell| \ge b|\delta_\ell|} \,\big] \\
&= \int_{\mathbb{R}} \int_{-2^{-\ell\beta/2}|z|}^{2^{-\ell\beta/2}|z|} \rho_\ell(y,z) \mathrm{d}y \mathrm{d}z \\
&\le \int_{\mathbb{R}} \int_{-2^{-\ell\beta/2}|z|}^{2^{-\ell\beta/2}|z|} p_{0,\ell}(z) \mathrm{d}y \mathrm{d}z \\
&\le 2 \times 2^{-\ell\beta/2} \int_{\mathbb{R}} |z| p_{0,\ell}(z) \mathrm{d}z \\
&\lesssim 2^{-\ell\beta/2},
\end{aligned}$$

where we use Assumption 2.6 to bound $\int_{\mathbb{R}} |z| p_{0,\ell}(z)\mathrm{d}z$ uniformly in ℓ. ∎

The stricter conditions also give a tighter bound on the E_ℓ than Remark 2.4, and hence better MLMC complexity when $\beta < 2\gamma$.

PROPOSITION 2.9 ([18, Proposition 1]). *Let Assumptions 2.6 and 2.7 hold for some $\beta > 0$, $\frac{\beta}{2} \le \alpha \le \beta$. Then, $|\mathbb{E}[\, \mathbb{H}(g) - \mathbb{H}(g_\ell) \,]| \lesssim 2^{-\alpha\ell}$.*

Proof. For $\rho_\ell(y,z)$ given by Assumption 2.6 we have

$$\mathbb{E}[\, \mathbb{H}(g) \,] = \int_{\mathbb{R}} \int_{2^{-\beta\ell/2} z}^{\infty} \rho_\ell(y,z) \mathrm{d}y \mathrm{d}z.$$

Thus

$$\begin{aligned}
|\mathbb{E}[\, \mathbb{H}(g) - \mathbb{H}(g_\ell) \,]| &= |\mathbb{E}[\, \mathbb{H}(g_\ell) \,] - \mathbb{E}[\, \mathbb{H}(g) \,]| \\
&= \left| \int_{\mathbb{R}} \int_0^{2^{-\beta\ell/2}} \rho_\ell(y,z) \mathrm{d}y \mathrm{d}z \right|.
\end{aligned}$$

A Taylor expansion gives

$$\rho_\ell(y,z) = \rho_\ell(0,z) + y \frac{\partial}{\partial y} \rho_\ell(0,z) + \frac{y^2}{2} \frac{\partial^2}{\partial y^2} \rho_\ell(\hat{y}, z), \tag{2.5}$$

for some $\hat{y} \in [0, y]$. Inserting this into the double integral above and using Assumptions 2.6 and 2.7 gives

$$\begin{aligned} E_\ell &\leq \left| 2^{-\beta\ell/2} \int_{\mathbb{R}} z \rho_\ell(0,z) \mathrm{d}z \right| + 2^{-\beta\ell} \int_{\mathbb{R}} |z|^2 p_{1,\ell}(z) \mathrm{d}z \\ &\qquad + 2^{-3\beta\ell/2} \int_{\mathbb{R}} |z|^3 p_{2,\ell}(z) \mathrm{d}z \\ &\lesssim 2^{-\beta\ell/2} \left| \mathbb{E}[\, Z_\ell \mid \delta_\ell = 0 \,] \right| + \mathcal{O}\big(2^{-\beta\ell}\big) \\ &\lesssim 2^{-\alpha\ell}, \end{aligned}$$

where we used Assumption 2.7 and the definition of Z_ℓ to bound $\mathbb{E}[\, Z_\ell \mid \delta_\ell = 0 \,] \lesssim 2^{\ell(\beta/2-\alpha)}$ and assume $\int_{\mathbb{R}} \rho_\ell(0,z)\mathrm{d}z > 0$ as in the proof of Proposition 2.8. ∎

The discussion above proves the following complexity results.

COROLLARY 2.10. *Under Assumptions* 1.3 *and* 1.4*, the total work required for the MLMC estimator* (2.1) *with mean square error* ε^2 *can be bounded by*

$$\mathrm{Work}(\mathcal{M}^*; \varepsilon) \lesssim \begin{cases} \varepsilon^{-2} & \beta > 2(\frac{q+1}{q})\gamma \\ \varepsilon^{-2}(\log \varepsilon)^2 & \beta = 2(\frac{q+1}{q})\gamma \\ \varepsilon^{-1-2(\frac{q+1}{q})\gamma/\beta} & \beta < 2(\frac{q+1}{q})\gamma \end{cases}.$$

Proof. The result follows by combining Proposition 2.3 and Remark 2.4 with Proposition 2.1 for $\Delta\mathbb{H}_\ell = \mathbb{H}(g_\ell) - \mathbb{H}(g_{\ell-1})$. ∎

COROLLARY 2.11. *Under Assumption* 2.6 *and, when* $\beta < 2\gamma$*, also under Assumption* 2.7 *the total work required for the MLMC estimator* (2.1) *with mean square error* ε^2 *can be bounded by*

$$\mathrm{Work}(\mathcal{M}^*; \varepsilon) \lesssim \begin{cases} \varepsilon^{-2} & \beta > 2\gamma \\ \varepsilon^{-2}(\log \varepsilon)^2 & \beta = 2\gamma \\ \varepsilon^{-2-(\gamma-\beta/2)/\alpha} & \beta < 2\gamma \end{cases}$$

Proof. The result follows by combining Proposition 2.3 and Proposition 2.9 with Proposition 2.1 for $\Delta\mathbb{H}_\ell = \mathbb{H}(g_\ell) - \mathbb{H}(g_{\ell-1})$. ∎

In some applications, Assumption 1.3 holds for all $q < \infty$. Under the conditions of Corollary 2.10 and by considering arbitrarily large values of q we then bound the total work by

$$\mathrm{Work}(\mathcal{M}^*; \varepsilon) \lesssim \begin{cases} \varepsilon^{-2} & \beta > 2\gamma \\ \varepsilon^{-1-\nu-2\gamma/\beta} & \beta \leq 2\gamma \end{cases},$$

for any $\nu > 0$.

For Examples 1.1 and 1.2 with Euler-Maruyama simulation of the SDE, we can show (under certain assumptions on the underlying SDE [21]) that $\alpha = \beta = \gamma$ and the complexity is at best $\varepsilon^{-5/2}$, a significant increase over the canonical ε^{-2} complexity. For SDE simulation we can replace the $\varepsilon^{-\nu}$ term appearing in the complexity in the limit $q \to \infty$ with a logarithmic factor using the analysis in [1].

3. Adaptive Multilevel Monte Carlo.

In the previous section, we described how the complexity of MLMC calculations for the problem (1.2) is affected by the discontinuous observable $\mathbb{H}(g)$. To improve the performance of MLMC we replace the approximation g_ℓ at level ℓ with $g_{\ell+\eta_\ell}$. Where we introduce the random, non-negative, integer η_ℓ which should reflect the uncertainty in the sign of $g_{\ell+\eta_\ell}$. The MLMC estimator (2.4) then uses the multilevel correction term $\Delta\mathbb{H}_\ell$ given by

$$\Delta\mathbb{H}_\ell \coloneqq \begin{cases} \mathbb{H}(g_{\ell+\eta_\ell}) - \mathbb{H}(g_{\ell-1+\eta_{\ell-1}}) & \ell > 0 \\ \mathbb{H}(g_{\eta_0}) & \ell = 0 \end{cases} \tag{3.1}$$

Heuristically, approximations which are close to zero with high variability should be refined further (have larger values of η_ℓ) than approximations which lie far away from zero with low variability. The chosen approach for sampling $g_{\ell+\eta_\ell}$ is detailed in Algorithm 3.1. We refine between levels $\ell \leq \ell + \eta_\ell \leq \ell + \lceil \theta\ell \rceil$, for a supplied parameter θ, based on the value of $|\delta_{\ell+\eta_\ell}|$ (1.6). Algorithm 3.1 also has the parameter r, determining how strict we are with the refinement, and a confidence constant $c > 0$. Algorithm 3.1 implies that we refine by η_ℓ levels, where

$$\eta_\ell = k \iff \begin{cases} |\delta_{\ell+m}| < c2^{\gamma(\theta\ell(1-r)-m)/r} & \forall m \leq k-1 \\ |\delta_{\ell+k}| \geq c2^{\gamma(\theta\ell(1-r)-k)/r} & \text{if } k < \theta\ell \end{cases}, \tag{3.2}$$

for $0 \leq k \leq \lceil \theta\ell \rceil$. For small values of r we refine samples to higher levels than for large r. Ideally, we want to allow the refinement procedure to take r as large as possible while observing maximum benefit to the MLMC complexity. For the MLMC computation to converge to the correct mean, it is important that the method of refining $g_{\ell+k}$ to $g_{\ell+k+1}$ does not affect the almost sure convergence of $g_{\ell+\eta_\ell}$ to g. To ensure the cost of computing $\sigma_{\ell+k}$ does not dominate the refinement, we assume throughout that the cost of computing $\sigma_{\ell+k}$ is of order $2^{\gamma(\ell+k)}$.

EXAMPLE 3.1. *For the nested simulation problem (Example* 1.1*), we can refine $g_{\ell+k}$ to $g_{\ell+k+1}$ by sampling an additional $N_{\ell+k+1} - N_{\ell+k}$ samples of X given Y to use in the refined Monte Carlo average. Alternatively, we may sample $N_{\ell+k+1}$ new, independent samples of X given the same value of Y to compute $g_{\ell+k+1}$. $\sigma_{\ell+k}$ may be computed as the sample standard deviation of $N_{\ell+k}$ samples.*

EXAMPLE 3.2. *For digital option pricing (Example* 1.2*), the underlying Brownian path of the SDE can be refined using the Brownian Bridge construction, and $\sigma_{\ell+k}$ can be chosen to be a constant. See Subsection* 4.2 *for more details.*

Algorithm 3.1 Adaptive sampling at level ℓ

Input: $\ell, r, \theta, c > 0, \gamma, \beta$
Output: Adaptively refined sample $g_{\ell+\eta_\ell}$
Set $k = 0$
Sample (g_ℓ, σ_ℓ)
Compute δ_ℓ given (g_ℓ, σ_ℓ)
while $|\delta_{\ell+k}| < c2^{\gamma(\theta\ell(1-r)-k)/r}$ and $k < \lceil \theta\ell \rceil$ **do**
  Refine $(g_{\ell+k}, \sigma_{\ell+k})$ to $(g_{\ell+k+1}, \sigma_{\ell+k+1})$
  Compute $\delta_{\ell+k+1}$ given $(g_{\ell+k+1}, \sigma_{\ell+k+1})$
  Set $k = k + 1$
end while
Set $\eta_\ell = k$
return $g_{\ell+\eta_\ell}$

Algorithm 3.1 has many similarities to the adaptive nested simulation algorithm in [13, 14], which considers the specific case $g = \mathbb{E}[\,X|Y\,]$ approximated by an inner

Monte Carlo sampler. However, besides being applicable to a wider class of problems, the present algorithm has some key differences: The nested simulation algorithm in [13, 14] requires that each refined value $g_{\ell+k+1}$ is independent of the previous term $g_{\ell+k}$ conditioned on Y, which is not required here. This accelerates the refinement procedure since one can reuse all terms from the computation of $g_{\ell+k}$ in the refinement to $g_{\ell+k+1}$. Moreover, in [13] the adaptive algorithm returns only the number of inner samples one should use to approximate $\mathbb{E}[\,X|Y\,]$, given Y, and the estimate of g should then be computed independently. In contrast, Algorithm 3.1 requires that the estimate of g matches the output of the refinement process. The parameter θ is also a novel introduction to Algorithm 3.1. In [13], the nested simulation application has $\beta = \gamma$ for which the value $\theta = 1$ is optimal (see Lemma 3.7). For $\beta \neq \gamma$ it can be optimal to refine over a wider or narrower range of levels, see Lemma 3.4 and Remark 3.5. In [13], the theory requires the stronger assumption that

$$\sup_y \mathbb{E}\Big[\, \mathrm{Var}[\,X|Y\,]^{-q/2} |X - \mathbb{E}[\,X|Y\,]|^q \;\Big|\; Y = y \Big] < \infty,$$

for some $2 < q < \infty$ which results in a different analysis to that presented below. However, for most practical examples the key results are similar.

To satisfy (2.2) and (2.3), one must typically correlate the fine and coarse components of $\Delta\mathbb{H}_\ell$. The adaptivity introduced in Algorithm 3.1 does not impact this correlation. In fact, the algorithm can be modified naturally to compute $g_{\ell+\eta_\ell}$ and $g_{\ell-1+\eta_{\ell-1}}$ simultaneously using correlated noise, provided that $\eta_{\ell-1}$ is also chosen according to Algorithm 3.1 with $\ell-1 \leftarrow \ell$. Since η_ℓ and $\eta_{\ell-1}$ are both chosen according to Algorithm 3.1, it is worth noting whether it is possible that in some circumstances the 'coarse' estimator $g_{\ell-1+\eta_{\ell-1}}$ is in fact refined further by Algorithm 3.1 than the 'fine' estimator $g_{\ell+\eta_\ell}$. We consider the case where the refinement of the fine estimator $g_{\ell+\eta_\ell}$ is correlated to that of the coarse estimator $g_{\ell-1+\eta_{\ell-1}}$ such that when $\eta_{\ell-1} = \eta_\ell + 1$ we have $\delta_{\ell+\eta_\ell} = \delta_{\ell-1+\eta_{\ell-1}}$. For example, this is false in the nested simulation problem discussed in Example 3.1 if one uses independent samples of X given Y for the fine and coarse estimator, but is true for the digital option problem considered in Example 3.2 when the fine and coarse estimator use the same underlying Brownian path. In this case, when $r \leq \theta^{-1}+1$ it follows from (3.2) that $\eta_{\ell-1} \leq \eta_\ell+1$. However, when $r > \theta^{-1}+1$ there is a small chance that the 'coarse' sample, $g_{\ell-1+\eta_{\ell-1}}$, is actually refined to greater accuracy than the 'fine' estimator, $g_{\ell+\eta_\ell}$. Proposition 3.3 below assures that on average $g_{\ell+\eta_\ell}$ has greater accuracy than $g_{\ell-1+\eta_{\ell-1}}$.

3.1. Work Analysis. In the context of Proposition 2.1, using $\Delta\mathbb{H}_\ell$ as in (3.1) we wish to improve upon the convergence rate of V_ℓ seen for the estimator (2.1) in Proposition 2.3. Proposition 2.1 implies that for this to be effective the expected cost of computing $g_{\ell+\eta_\ell}$ and g_ℓ must be similar. The following result ensures the expected cost of sampling $g_{\ell+\eta_\ell}$ is also $\mathcal{O}(2^{\gamma\ell})$.

PROPOSITION 3.3 ([13, Theorem 2.7]). *Define η_ℓ as in* (3.2) *and assume Assumption* 1.4 *holds for fixed $\rho_0, \delta > 0$. Provided $r > 1$, we have*

$$\mathbb{E}\Big[\, 2^{\gamma(\ell+\eta_\ell)} \Big] \lesssim 2^{\gamma\ell}.$$

Proof. We start with

$$\begin{aligned}\mathbb{E}[\, 2^{\gamma(\ell+\eta_\ell)} \,] &= \sum_{k=0}^{\lceil \theta\ell \rceil} 2^{\gamma(\ell+k)} \mathbb{P}[\, \eta_\ell = k \,] \\ &\le 2^{\gamma\ell} + \sum_{k=1}^{\lceil \theta\ell \rceil} 2^{\gamma(\ell+k)} \mathbb{P}[\, |\delta_{\ell+k-1}| < c 2^{\gamma(\theta\ell(1-r)-k+1)/r} \,],\end{aligned}$$

where we used (3.2) to bound the probabilities. Provided $r > 1$, for large enough ℓ we have $c2^{\gamma(\theta\ell(1-r)-k+1)/r} < \delta$ for all $k \ge 0$. Using Assumption 1.4

$$\begin{aligned}\mathbb{E}[\, 2^{\gamma(\ell+\eta_\ell)} \,] &\le 2^{\gamma\ell} + \rho_0 c 2^{\gamma/r} 2^{\gamma\ell(1+\theta(1-r)/r)} \sum_{k=1}^{\lceil \theta\ell \rceil} 2^{\gamma k(r-1)/r} \\ &\le 2^{\gamma\ell} + c_0 2^{\gamma\ell} 2^{\gamma\ell\theta(1-r)/r} 2^{\gamma\lceil \theta\ell \rceil (r-1)/r} \\ &\lesssim 2^{\gamma\ell},\end{aligned}$$

since $r > 1$. ∎

Note that the above proof emphasises that a larger r value results in a lower sampling cost.

3.2. Analysis of the Variance. The following results highlight improvements to the convergence of $\mathbb{E}[\, (\mathbb{H}(g_\ell) - \mathbb{H}(g_{\ell-1}))^2 \,]$ to 0 under the adaptive sampling procedure in Algorithm 3.1. As with the non-adaptive case, we obtain slightly better results using the stronger Assumption 2.6. However, this condition is not essential and we still see an improvement under the general Assumptions 1.3 and 1.4, as seen below.

LEMMA 3.4. *Let Assumptions* 1.3 *and* 1.4 *hold for some* $\beta > 0$ *and* $q > 2$. *Assume:*

- *For* $\beta \le (\frac{q+1}{q})\gamma$ *we take*

$$\theta = \left(2\left(\frac{q+1}{q}\right)\frac{\gamma}{\beta} - 1 \right)^{-1}, \tag{3.3}$$

  and $r < 2\frac{\gamma}{\beta}$.
- *For* $\beta > (\frac{q+1}{q})\gamma$ *we take* $\theta = 1$ *and*

$$\begin{cases} r \le \left(1 - \frac{(q-1)\beta}{2(q+1)\gamma}\right)^{-1} & \text{when } \beta < 2\left(\frac{q+1}{q-1}\right)\gamma \\ r < \infty & \text{when } \beta \ge 2\left(\frac{q+1}{q-1}\right)\gamma \end{cases}. \tag{3.4}$$

Then, for $g_{\ell+\eta_\ell}$ *given by Algorithm* 3.1,

$$\mathbb{E}[\, (\mathbb{H}(g) - \mathbb{H}(g_{\ell+\eta_\ell}))^2 \,] \lesssim 2^{-(\frac{q}{q+1})\ell\beta(1+\theta)/2}. \tag{3.5}$$

Proof. As with the work analysis, we split the calculation across possible values of η_ℓ

$$
\begin{aligned}
\mathbb{E}[\,(\mathbb{H}(g)-\mathbb{H}(g_{\ell+\eta_\ell}))^2\,] &= \sum_{k=0}^{\lceil\theta\ell\rceil}\mathbb{E}[\,(\mathbb{H}(g)-\mathbb{H}(g_{\ell+\eta_\ell}))^2\mathbb{I}_{\eta_\ell=k}\,]\\
&\le \underbrace{\sum_{k=0}^{\lceil\theta\ell\rceil-1}\mathbb{E}[\,(\mathbb{H}(g)-\mathbb{H}(g_{\ell+k}))^2\mathbb{I}_{\eta_\ell=k}\,]}_{=:\Sigma_0}\\
&\quad+\underbrace{\mathbb{E}[\,(\mathbb{H}(g)-\mathbb{H}(g_{\ell+\lceil\theta\ell\rceil}))^2\,]}_{=:\Sigma_1}
\end{aligned}
\tag{3.6}
$$

By Proposition 2.3 we have

$$\Sigma_1 \lesssim 2^{-(\frac{q}{q+1})\ell\beta(1+\theta)/2}.$$

We now turn our attention to terms for which $k<\lceil\theta\ell\rceil$. Using (3.2) to relate the condition $\eta_\ell=k$ to the value of $\delta_{\ell+k}$ we have

$$
\begin{aligned}
\Sigma_0 &\le \sum_{k=0}^{\lceil\theta\ell\rceil-1}\mathbb{E}[\,\mathbb{I}_{|g-g_{\ell+k}|\ge|g_{\ell+k}|}\mathbb{I}_{\eta_\ell=k}\,]\\
&\le \sum_{k=0}^{\lceil\theta\ell\rceil-1}\mathbb{E}\Big[\,\mathbb{I}_{|g-g_{\ell+k}|\ge|g_{\ell+k}|}\quad \mathbb{I}_{|\delta_{\ell+k}|\ge c2^{(\theta\ell(1-r)-k)\gamma/r}}\,\Big]\\
&\le \sum_{k=0}^{\lceil\theta\ell\rceil-1}\mathbb{E}\Big[\,\mathbb{I}_{a_k\le|Z_{\ell+k}|b_k^{-1}}\,\Big]
\end{aligned}
$$

where $Z_{\ell+k}$ is as in (1.4) and we introduce the terms

$$
\begin{aligned}
a_k &:= c\,2^{(\theta\ell(1-r)-k)\gamma/r}\\
b_k &:= 2^{(\ell+k)\beta/2}.
\end{aligned}
$$

Using Assumption 1.3 and Markov's inequality (1.5) we obtain

$$\Sigma_0 \le \sum_{k=0}^{\lceil\theta\ell\rceil-1} a_k^{-q}b_k^{-q}\mathbb{E}[\,|Z_{\ell+k}|^q\,].$$

Thus we restrict our attention to the term

$$a_k^{-q}b_k^{-q} = 2^{-q(\beta/2-\gamma/r)(k+\ell)}2^{-q\gamma\ell(1+\theta-\theta r)/r}. \tag{3.7}$$

Suppose first that $\beta\le(\frac{q+1}{q})\gamma$. Using the assumption that $r<2\frac{\gamma}{\beta}$ in this case, (3.7) is an increasing function of k. It follows that

$$
\begin{aligned}
\sum_{k=0}^{\lceil\theta\ell\rceil-1} a_k^{-q}b_k^{-q}\mathbb{E}[\,|Z_{\ell+k}|^q\,] &\lesssim a_{\theta\ell}^{-q}b_{\theta\ell}^{-q}\\
&\lesssim 2^{q\ell(\gamma\theta-\beta(1+\theta)/2)}.
\end{aligned}
\tag{3.8}
$$

In order to ensure the above term is of the same order as Σ_1 we take θ as in (3.3).

Now suppose $\beta > (\frac{q+1}{q})\gamma$ and consider first $r < 2\frac{\gamma}{\beta}$ so that (3.8) holds. Note that taking $\theta = 1$ is enough to guarantee $\Sigma_1 \lesssim 2^{-(\frac{q}{q+1})\beta\ell}$ and, by (3.8), $\Sigma_0 \lesssim 2^{q\ell(\gamma-\beta)} \leq 2^{-(\frac{q}{q+1})\beta\ell}$ since $\beta \geq (\frac{q+1}{q})\gamma$. Since $\beta > (\frac{q+1}{q})\gamma$ this is enough to guarantee ε^{-2} complexity. If $r = 2\frac{\gamma}{\beta}$ the bound (3.8) becomes (again taking $\theta = 1$)

$$
\begin{aligned}
\Sigma_0 \leq \sum_{k=0}^{\ell-1} a_k^{-q} b_k^{-q} \mathbb{E}[\,|Z_{\ell+k}|^q\,] &= \sum_{k=0}^{\ell-1} 2^{q\ell(\gamma-\beta)} \\
&\lesssim \ell 2^{q\ell(\gamma-\beta)} \lesssim 2^{-(\frac{q}{q+1})\beta\ell}.
\end{aligned}
$$

On the other hand, for $r > \frac{2\gamma}{\beta}$, (3.7) is a decreasing function of k and we have

$$
\begin{aligned}
\Sigma_0 \leq \sum_{k=0}^{\ell-1} a_k^{-q} b_k^{-q} \mathbb{E}[\,|Z_{\ell+k}|^q\,] &\lesssim a_0^{-q} b_0^{-q} \\
&\lesssim 2^{q(\gamma(r-1)/r-\beta/2)\ell} \lesssim 2^{-(\frac{q}{q+1})\beta\ell},
\end{aligned}
\qquad \square
$$

provided we take r as in (3.4), completing the proof.

REMARK 3.5. *The proof of Lemma* 3.4 *allows* (3.5) *to hold for certain values* $\theta > 1$ *provided* $\beta > (\frac{q+1}{q})\gamma$ *and under tighter upper bounds for* r*. However, for such values of* β *we are already in the* ε^{-2} *complexity regime of MLMC at* $\theta = 1$*, thus any increase in* θ *can improve the MLMC cost by a constant at best. Moreover, tighter bounds on* r *will increase the expected cost of sampling* $g_{\ell+\eta_\ell}$*, limiting the value of any constant reduction in the MLMC cost.*

REMARK 3.6. *As with the non-adaptive MLMC, the absolute moments of* $\mathbb{H}(g_\ell) - \mathbb{H}(g_{\ell-1})$ *are equal, thus Lemma* 3.4 *actually proves a bound for all even moments of* $\mathbb{H}(g_\ell) - \mathbb{H}(g_{\ell-1})$.

Below, we state an extension to Lemma 3.4 under the stricter assumptions required for the bias analysis.

LEMMA 3.7. *Let Assumption* 2.6 *hold for some* $\beta > 0$ *and* $q > 2$*. Assume:*

- *For* $\beta \leq \gamma$ *we take*

$$
\begin{aligned}
\theta &= \left(2\frac{\gamma}{\beta} - 1\right)^{-1}, \\
r &< 2\frac{\gamma}{\beta}\left(1 - \frac{1}{q}\right).
\end{aligned}
\tag{3.9}
$$

- *For* $\beta > \gamma$ *we take* $\theta = 1$ *and*

$$
\begin{cases}
r \leq \left(1 - \frac{(q-2)\beta}{2(q-1)\gamma}\right)^{-1} & \text{when } \beta < 2\left(\frac{q-1}{q-2}\right)\gamma \\
r < \infty & \text{when } \beta \geq 2\left(\frac{q-1}{q-2}\right)\gamma
\end{cases}.
\tag{3.10}
$$

Then, for $g_{\ell+\eta_\ell}$ *as in Algorithm* 3.1

$$
\mathbb{E}[\,(\mathbb{H}(g) - \mathbb{H}(g_{\ell+\eta_\ell}))^2\,] \lesssim 2^{-\beta\ell\cdot(1+\theta)/2}.
\tag{3.11}
$$

Proof. As in the previous result, we split the calculation across all refined levels as in (3.6). By Proposition 2.8, it follows that $\Sigma_1 \lesssim 2^{-\beta(1+\theta)\ell/2}$. Moreover, for $k < \lceil \theta\ell \rceil$ and defining a_k, b_k as in the proof of Lemma 3.4 we have

$$\begin{aligned}
&\mathbb{E}\Big[(\mathbb{H}(g) - \mathbb{H}(g_{\ell+k}))^2 \mathbb{I}_{\eta_\ell = k} \Big] \\
&= \mathbb{E}\Big[\Big(\mathbb{I}_{0 > \delta_{\ell+k} > b_k^{-1} Z_{\ell+k}} + \mathbb{I}_{b_k^{-1} Z_{\ell+k} > \delta_{\ell+k} > 0} \Big) \mathbb{I}_{|\delta_{\ell+k}| \geq c \cdot a_k} \Big] \\
&= \mathbb{E}\Big[\Big(\mathbb{I}_{0 > \delta_{\ell+k} > b_k^{-1} Z_{\ell+k}} + \mathbb{I}_{b_k^{-1} Z_{\ell+k} > \delta_{\ell+k} > 0} \Big) \mathbb{I}_{b_k^{-1}|Z_{\ell+k}| \geq |\delta_{\ell+k}| \geq c \cdot a_k} \Big] \\
&\leq a_k^{1-q} b_k^{1-q} \mathbb{E}\Big[|Z_{\ell+k}|^{q-1} \Big(\mathbb{I}_{0 > \delta_{\ell+k} > b_k^{-1} Z_{\ell+k}} + \mathbb{I}_{b_k^{-1} Z_{\ell+k} > \delta_{\ell+k} > 0} \Big) \Big] \\
&= a_k^{1-q} b_k^{1-q} \Big(\int_0^\infty \int_0^{b_k^{-1} z} |z|^{q-1} \rho_{\ell+k}(y,z) \mathrm{d}y \mathrm{d}z + \int_{-\infty}^0 \int_{-b_k^{-1} z}^0 |z|^{q-1} \rho_{\ell+k}(y,z) \mathrm{d}y \mathrm{d}z \Big).
\end{aligned}$$

By Assumption 2.6 we can bound $\rho_{\ell+k}(y,z)$ from above by $p_{0,\ell+k}(z)$ and obtain

$$\begin{aligned}
\mathbb{E}\Big[(\mathbb{H}(g) - \mathbb{H}(g_{\ell+k}))^2 \mathbb{I}_{\eta_\ell = k} \Big] &\lesssim a_k^{1-q} b_k^{-q} \\
&= 2^{\frac{\gamma}{r}(\ell + \theta\ell(1-r))(1-q)} 2^{((q-1)\gamma/r - q\beta/2)(\ell+k)}.
\end{aligned}$$

When $r < \frac{2\gamma}{\beta}(1 - 1/q)$, this above term is dominant when $k = \lceil \theta\ell \rceil$. It follows that one can make the orders of Σ_0 and Σ_1 equal as $\ell \to \infty$ in (3.6) by taking θ as in (3.9). When $\beta \geq \gamma$, instead we fix $\theta = 1$. A similar calculation to Lemma 3.4 then shows the result holds provided r satisfies (3.10). ∎

3.3. Analysis of the Bias.

In the context of Lemma 3.7, Proposition 2.1 implies the complexity of (adaptive) MLMC is affected by the convergence rate of $E_\ell = |\mathbb{E}[\,\mathbb{H}(g) - \mathbb{H}(g_{\ell+\eta_\ell})\,]|$ whenever $\beta < \gamma$. To improve the rate given by Proposition 2.9 due to adaptive sampling, we make a further assumption.

Assumption 3.8. *Define $Z_\ell, \beta > 0$ as in Assumption 1.3 and $\frac{\beta}{2} \leq \alpha \leq \beta$ as in Assumption 2.7. Then, for $j = 0, 1$ and all $\ell \in \mathbb{N}, x \geq 0$ we have*

$$\left| \mathbb{E}\left[\operatorname{sign}(Z_\ell) |Z_\ell|^j \;\middle|\; |Z_\ell| \geq x \right] \right| \lesssim 2^{\ell(\beta/2 - \alpha)}.$$

By Assumptions 2.6 and 2.7 we know that this condition holds for $j = 1$ and $x = 0$. Assumption 3.8 ensures that the mean of Z_ℓ converges at the same rate even when conditioned on taking large values. When $j = 0$ the assumption implies that the probability of observing large positive Z_ℓ is reasonably close to the probability of observing large negative Z_ℓ. The necessity for this assumption arises since the refined samples are only accepted before the maximum level if $|\delta_\ell|$ is sufficiently large. As such, the error $\mathbb{H}(g) - \mathbb{H}(g_\ell)$ is non-zero only for suitably large values of Z_ℓ. The resulting improvement to E_ℓ is discussed below.

Lemma 3.9. *Let Assumptions 2.6 and 3.8 hold for $\beta > 0$ and $\frac{\beta}{2} \leq \alpha \leq \beta$. For $\beta \leq \gamma$, if we tighten the bound on r in Lemma 3.7 to $r < 2\frac{\gamma}{\beta}\left(\frac{q-2}{q}\right)$, then for $g_{\ell+\eta_\ell}$ as in Algorithm 3.1 and θ as in (3.9) we have*

$$|\mathbb{E}[\,\mathbb{H}(g) - \mathbb{H}(g_{\ell+\eta_\ell})\,]| \lesssim 2^{-\alpha(1+\theta)\ell}.$$

Proof. We bound

$$|\mathbb{E}[\,\mathbb{H}(g) - \mathbb{H}(g_{\ell+\eta_\ell})\,]| \le \sum_{k=0}^{\lceil\theta\ell\rceil-1} |\mathbb{E}[\,(\mathbb{H}(g) - \mathbb{H}(g_{\ell+k}))\mathbb{I}_{\eta_\ell=k}\,]| + \left|\mathbb{E}\left[\,\mathbb{H}(g) - \mathbb{H}(g_{\ell+\lceil\theta\ell\rceil})\,\right]\right|.$$

By Proposition 2.9 we know that the final term satisfies

$$\left|\mathbb{E}\left[\,\mathbb{H}(g) - \mathbb{H}(g_{\ell+\lceil\theta\ell\rceil})\,\right]\right| \lesssim 2^{-\alpha(1+\theta)\ell}. \tag{3.12}$$

By expanding the difference $\mathbb{H}(g)-\mathbb{H}(g_{\ell+k})$ according to when the difference is either ± 1 and considering the event $\eta_\ell = k$ we arrive at

$$\begin{aligned}
&|\mathbb{E}[\,(\mathbb{H}(g) - \mathbb{H}(g_{\ell+k}))\mathbb{I}_{\eta_\ell=k}\,]| \\
&= \left|\mathbb{E}\left[\,\left(\mathbb{I}_{b_k^{-1}Z_{\ell+k}<\delta_{\ell+k}<0} - \mathbb{I}_{0<\delta_{\ell+k}<b_k^{-1}Z_{\ell+k}}\right)\mathbb{I}_{|\delta_{\ell+k}|\ge c\cdot a_k}\,\right]\right| \\
&= \left|\mathbb{E}\left[\,\left(\mathbb{I}_{b_k^{-1}Z_{\ell+k}<\delta_{\ell+k}<a_k} - \mathbb{I}_{a_k<\delta_{\ell+k}<b_k^{-1}Z_{\ell+k}}\right)\,\right]\right| \\
&= \left|\int_{-\infty}^{-b_k a_k}\int_{b_k^{-1}z}^{-a_k} \rho_{\ell+k}(y,z)\mathrm{d}y\mathrm{d}z - \int_{b_k a_k}^{\infty}\int_{a_k}^{b_k^{-1}z} \rho_{\ell+k}(y,z)\mathrm{d}y\mathrm{d}z\right|,
\end{aligned}$$

where a_k, b_k are as in the proof of Lemma 3.4. We again use the Taylor expansion (2.5) on the density $\rho_{\ell+k}(y,z)$. The absolute value of the zero'th-order term is

$$\begin{aligned}
&\left|\int_{-\infty}^{-b_k a_k}\big(-a_k - b_k^{-1}z\big)\rho_{\ell+k}(0,z)\mathrm{d}z + \int_{b_k a_k}^{\infty}\big(a_k - b_k^{-1}z\big)\rho_{\ell+k}(0,z)\mathrm{d}z\right| \\
&\le a_k\left|\mathbb{E}\left[\,\mathrm{sign}(Z_{\ell+k})\mathbb{I}_{|Z_{\ell+k}|\ge a_k b_k}|\delta_{\ell+k}=0\,\right]\right| + b_k^{-1}|\mathbb{E}\left[\,Z_{\ell+k}\mathbb{I}_{|Z_{\ell+k}|\ge b_k a_k}|\delta_{\ell+k}=0\,\right]| \\
&\lesssim \mathbb{P}[\,|Z_{\ell+k}|\ge b_k a_k\,]\Bigg(a_k\left|\mathbb{E}\left[\,\mathrm{sign}(Z_{\ell+k})\;\middle|\;|Z_{\ell+k}|\ge b_k a_k, \delta_{\ell+k}=0\,\right]\right| \\
&\qquad + b_k^{-1}\left|\mathbb{E}\left[\,Z_{\ell+k}\;\middle|\;|Z_{\ell+k}|\ge b_k a_k, \delta_{\ell+k}=0\,\right]\right|\Bigg) \\
&\lesssim a_k^{1-q}b_k^{-q}2^{(\ell+k)(\beta/2-\alpha)}
\end{aligned}$$

where we used Assumption 3.8 and bounded $\mathbb{P}[\,|Z_{\ell+k}|\ge b_k a_k\,] \le a_k^{-q}b_k^{-q}\mathbb{E}[\,|Z_k|^q\,]$. For the first-order term, we obtain

$$\begin{aligned}
&\left|a_k^2\int_{-\infty}^{\infty}\mathbb{I}_{|z|\ge b_k a_k}\frac{\partial}{\partial y}\rho_{\ell+k}(0,z)\mathrm{d}z - b_k^{-2}\int_{-\infty}^{\infty}\mathbb{I}_{|z|\ge b_k a_k}z^2\frac{\partial}{\partial y}\rho_{\ell+k}(0,z)\mathrm{d}z\right| \\
&\le \left|b_k^{-q}a_k^{2-q}\int_{-\infty}^{\infty}|z|^q\frac{\partial}{\partial y}\rho_{\ell+k}(0,z)\mathrm{d}z - b_k^{-2-q}a_k^{-q}\int_{-\infty}^{\infty}|z|^{2+q}\frac{\partial}{\partial y}\rho_{\ell+k}(0,z)\mathrm{d}z\right| \\
&\lesssim b_k^{-q}a_k^{2-q}
\end{aligned}$$

by Assumption 2.6. Similarly, we can bound the second-order term up to a constant by $a_k^{3-q}b_k^{-q}$. Consequently, we have

$$\sum_{k=0}^{\lceil\theta\ell\rceil-1}|\mathbb{E}[\,(\mathbb{H}(g) - \mathbb{H}(g_{\ell+k}))\mathbb{I}_{\eta_\ell=k}\,]| \lesssim \sum_{k=0}^{\lceil\theta\ell\rceil-1} a_k^{1-q}b_k^{-q}2^{(\ell+k)(\beta/2-\alpha)} + \sum_{k=0}^{\lceil\theta\ell\rceil-1} a_k^{2-q}b_k^{-q}.$$

Provided $r < \frac{2\gamma}{\beta}\frac{q-2}{q}$, the dominant cost of each sum on the right hand side occurs at $k = \lceil\theta\ell\rceil - 1$, giving

$$\sum_{k=0}^{\lceil\theta\ell\rceil-1} |\mathbb{E}[\,(\mathbb{H}(g) - \mathbb{H}(g_{\ell+k}))\mathbb{I}_{\eta_\ell=k}\,]| \lesssim a_{\theta\ell}^{1-q} b_{\theta\ell}^{-q} 2^{\ell(1+\theta)(\beta/2-\alpha)} + a_{\theta\ell}^{2-q} b_{\theta\ell}^{-q}$$
$$\lesssim 2^{-\alpha(1+\theta)\ell} + 2^{-\beta(1+\theta)\ell},$$

for θ as in (3.9). ∎

Numerical tests suggest that the previous result does not hold when Assumption 3.8 is false, see Appendix A. However, one can still obtain reasonable convergence rates of E_ℓ without this result by Remark 2.4.

3.4. Bounds on $\mathrm{Work}(\mathcal{M}^*;\varepsilon)$**.** We conclude this section with a discussion on how the improved variance rate given by Lemma 3.4 affects the work bounds of MLMC. We begin by discussing the impact of adaptive sampling under the weaker assumptions.

THEOREM 3.10. *Under the assumptions of Lemma* 3.4*, the total work of MLMC using adaptive sampling as in Algorithm* 3.1 *with* $\Delta\mathbb{H}_\ell$ *given by* (3.1) *is*

$$\mathrm{Work}(\mathcal{M}^*;\varepsilon) \lesssim \begin{cases} \varepsilon^{-2} & \beta > (\frac{q+1}{q})\gamma \\ \varepsilon^{-2}(\log\varepsilon)^2 & \beta = (\frac{q+1}{q})\gamma \\ \varepsilon^{-2(\frac{q+1}{q})\gamma/\beta} & \beta < (\frac{q+1}{q})\gamma \end{cases}.$$

Proof. The result follows from applying Proposition 3.3, Lemma 3.4 to Proposition 2.1, with $\Delta\mathbb{H}_\ell$ given by (3.1). Similar to Remark 2.4 we can bound E_ℓ using Lemma 3.4. ∎

This result should be contrasted with Corollary 2.10. In particular, note how the canonical ε^{-2} complexity is obtained when $\beta > (\frac{q+1}{q})\gamma$ as opposed to when $\beta > 2(\frac{q+1}{q})\gamma$ for non-adaptive sampling. Moreover, even in the sub-optimal case when $\beta < (\frac{q+1}{q})\gamma$ the complexity is improved by a factor of ε^{-1} over the non-adaptive case. Often Assumption 1.3 holds for arbitrary $q < \infty$, in which case one can remove the q-dependence in Theorem 3.10 by adding a factor $\varepsilon^{-\nu}$ for any $\nu > 0$ to the complexity whenever $\beta \le \gamma$. When the assumptions of Lemma 3.9 hold, we obtain a slightly stronger result.

THEOREM 3.11. *Under the assumptions of Lemma* 3.7 *and, when* $\beta < \gamma$*, under the additional assumptions of Lemma* 3.9*, the total work of MLMC using adaptive sampling as in Algorithm* 3.1 *with* $\Delta\mathbb{H}_\ell$ *given by* (3.1) *is*

$$\mathrm{Work}(\mathcal{M}^*;\varepsilon) \lesssim \begin{cases} \varepsilon^{-2} & \beta > \gamma \\ \varepsilon^{-2}(\log\varepsilon)^2 & \beta = \gamma \\ \varepsilon^{-2-(1-\beta/(2\gamma))(\gamma-\beta)/\alpha} & \beta < \gamma \end{cases}.$$

Proof. The result follows by combining Proposition 3.3 and Lemma 3.7 with Proposition 2.1 for $\Delta\mathbb{H}_\ell$ given by (3.1). When $\beta < \gamma$ we use Lemma 3.9 to obtain a rate for E_ℓ. ∎

The previous result should be compared with Corollary 2.11. Again, we can see optimal complexities for β half as large as in the non-adaptive case. When $\beta < \gamma$ and $\alpha = \beta$ we can observe an improvement of order $\varepsilon^{-1/2}$ in the complexity.

4. Numerical Experiments. This section presents several numerical experiments to highlight the preceding theory[1]. We begin with some remarks on the technical components of MLMC.

Optimal Starting Level. In Section 2 we consider the MLMC estimator starting at level $\ell = 0$. When the approximations g_ℓ have pre-asymptotic behavior at small levels, it may be more efficient to start from some level $\ell_0 > 0$. For adaptive sampling, this is not the same as simply adjusting the work required at level $\ell = 0$ by a constant to account for a more accurate starting estimator. To see this, observe from Algorithm 3.1 that samples at level $\ell = 0$ cannot be refined further. In contrast, at level $\ell_0 > 0$ samples can be refined to maximum level $\ell_0 + \lceil \theta \ell_0 \rceil$. A heuristic approach for estimating the optimal starting level by a small computation is given in [13, Section 3]. We use optimal starting levels to obtain all MLMC estimates in the following sections.

Error Estimation. We illustrate the results of previous sections using the average work of sampling the multilevel correction term, W_ℓ, and the multilevel correction variance V_ℓ (2.3) and bias E_ℓ (2.2). Typically, V_ℓ and E_ℓ must be estimated using Monte Carlo sampling within MLMC. The robustness and accuracy of standard MLMC algorithms [11, 12] depends on reliable estimates of V_ℓ and E_ℓ to determine the optimal final level L and number of samples per level $\{M_\ell\}_{\ell_0 \le \ell \le L^*}$ required to have mean square error ε^2. Estimates of V_ℓ using a sample of size M_ℓ have standard deviation approximately given by $\sqrt{M_\ell^{-1}(\kappa_\ell - 1)}\mathrm{Var}[\, \Delta\mathbb{H}_\ell^2 \,]$, where κ_ℓ is the kurtosis of $\Delta\mathbb{H}_\ell$ [12, Section 3.3]. Thus we need $M_\ell \ge \kappa_\ell$ samples to obtain a reliable estimate for V_ℓ. From Remarks 2.5 and 3.6 it follows that $\kappa_\ell \approx V_\ell^{-1}$. Thus, we require more samples to reliably estimate V_ℓ as ℓ increases, which contradicts the intuition that MLMC aims to reduce the number of samples required at the finest levels. As a result, the robustness of MLMC can be affected by poor parameter estimation at the finest levels. One solution is detailed in [9], where E_ℓ and V_ℓ are approximated by Bayesian estimation with a beta prior distribution. An alternative solution, and the one used for the results stated here, is to estimate the proportionality constants in the bounds on V_ℓ and E_ℓ. We estimate these constants using the continuation MLMC approach discussed in [7].

4.1. Nested Expectation. The first numerical experiment is concerned with multilevel nested simulation [4, 13, 16, 24]. We take $g = \mathbb{E}[\, X|Y \,]$ so that (1.2) becomes $\mathbb{E}[\, \mathbb{H}(\mathbb{E}[\, X|Y \,]) \,]$. Approximations of g at a level ℓ are given by an inner Monte Carlo estimator

$$g_\ell = \frac{1}{N_\ell} \sum_{n=1}^{N_\ell} X^{(n)}(Y), \qquad X^{(n)}(Y) \overset{\text{i.i.d.}}{\sim} X|Y \tag{4.1}$$

using $N_\ell = N_0 2^{\gamma\ell}$ samples. When refining from level $\ell + k$ to $\ell + k + 1$ in Algorithm 3.1 we take the $N_{\ell+k}$ samples used to sample $g_{\ell+k}$ and add another $N_{\ell+k}(2^\gamma - 1)$ independent samples to form the sample of $g_{\ell+k+1}$. We assume σ_ℓ^2 is given by the sample variance of the samples used to generate g_ℓ, other choices of σ_ℓ are discussed in Appendix B. Using this choice of σ_ℓ, we can write Z_ℓ in (1.4) as

$$Z_\ell = \sqrt{\frac{N_0 N_\ell}{N_\ell - 1}} T_{N_\ell}, \tag{4.2}$$

[1] The code used for these experiments is written in Python, and can be found at https://github.com/JSpence97/mlmc-for-probabilities.

where T_{N_ℓ} is Student's t-statistic with samples $\{X^{(n)}(Y) - \mathbb{E}[\,X|Y\,]\}_{n=1}^{N_\ell}$. If the joint density $\hat{\rho}(x,y)$ of $X - \mathbb{E}[\,X|Y\,]$ and Y is bounded and monotone decreasing (increasing) for large positive (large negative) values of x, it follows from [20, Proposition 5.1, Theorem 6.2] that for each Y, $\mathbb{E}[\,|Z_\ell|^q \mid Y\,]$ is uniformly bounded in ℓ provided $N_\ell \geq q+1$. For reasonable X and Y we can extend this to a uniform bound in Y as well as ℓ, thus proving Assumption 1.3 for $q < N_\ell - 1$ using the Tower Property. By taking $q < N_\ell - 1$ we recover optimal results from the limit $q \to \infty$ only as $\ell \to \infty$. However, by taking a large enough number of inner samples at level 0, say $N_0 = 32$, we observe near asymptotic performance even at small levels.

Assumption 1.4 would follow by showing that $\delta_\infty \coloneqq |g|/\sqrt{\mathrm{Var}[\,X|Y\,]}$ has a density which is bounded in some open interval containing 0, as in [13]. We leave a more rigorous discussion of exact conditions required for the validity of these assumptions, and the effect of having $q < N_\ell - 1$ on the complexity of MLMC, to future work.

For comparison with [13] we consider the model problem used there, given by

$$X = \frac{2}{100}\left(Y^2 - Y_0^2\right) + \frac{7\sqrt{2}}{25} Y Y_1 - 0.0805$$

for $Y, Y_0, Y_1 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. For this problem, one has $\mathbb{E}[\,\mathbb{H}(\mathbb{E}[\,X|Y\,])\,] \approx 0.025$. In [13] the use of additional measures such as antithetic sampling of $\Delta\mathbb{H}_\ell$ [4,13] is considered to reduce the total cost of MLMC by a constant factor independent of the error bound ε. Such approaches can easily be altered to suit the present setup. We emphasize that the key difference between Algorithm 3.1 and the adaptive scheme in [13] for this setup is that here we do not re-sample all values of $X^{(n)}(Y)$ when refining to higher levels and the samples generated in Algorithm 3.1 are used to form our estimate of $g_{\ell+\eta_\ell}$, in contrast to [13,14].

The MLMC estimator is computed using non-adaptive sampling with $\gamma = 1, 2$ and adaptive sampling as in Algorithm 3.1 with $\gamma = 1$ and $r = 1.95, \theta = 1$ to fulfill the assumptions of Proposition 3.3 and Lemma 3.4 in the limit $q \to \infty$. The confidence constant is taken to be $c = 3/\sqrt{N_0}$, which aligns with the corresponding parameter in [13]. For each method, we plot W_ℓ, V_ℓ and E_ℓ versus ℓ.

Results are shown in Figure 4.1. The top left plot shows W_ℓ vs ℓ. By construction, the work per level for the non-adaptive schemes is a deterministic term proportional to $2^{\gamma\ell}$. For the adaptive scheme and $\ell > 2$, we observe $W_\ell \propto 2^\ell$, increased by a constant factor over the non-adaptive sampler with $\gamma = 1$. This agrees with Proposition 3.3, which states that adaptive sampling does not affect the rate at which W_ℓ increases. The variance V_ℓ per level is shown in the top right plot of Figure 4.1. Following from Proposition 2.3 with $q \to \infty$, the non-adaptive samplers have variance decreasing at rate $\beta/2 \approx \gamma/2$. Instead, the adaptive sampler matches the variance seen for the non-adaptive method with $\gamma = 2$, as predicted by Lemma 3.4. Moreover, in the bottom left plot of Figure 4.1, we see that the bias reduction rates guaranteed from Proposition 2.9 and Lemma 3.9 with $\alpha = \beta$. In other words, the adaptive scheme exhibits the same variance and bias reduction rate as the non-adaptive method with $\gamma = 2$, but has expected work per level comparable to the non-adaptive method with $\gamma = 1$.

In the bottom right plot of Figure 4.1, we display the total work of sampling $\mathcal{M}^*$ multiplied by ε^2 against the accuracy ε, normalised according to the true value 0.025. The total work is taken as the number of inner samples generated from X

for a given Y. For each method, we run the algorithm from an estimated optimal starting level as in [13]. The theoretical complexity rates given by Corollary 2.11 and Theorems 3.10 and 3.11 for $q \to \infty$ are plotted as dashed and dotted lines, highlighting the applicability of the preceding theory to this example. The adaptive approach is able to reduce the complexity from order $\varepsilon^{-5/2}$ to order $\varepsilon^{-2}(\log \varepsilon)^2$. This is equivalent to the performance in [13] and the observed numerical results are similar.

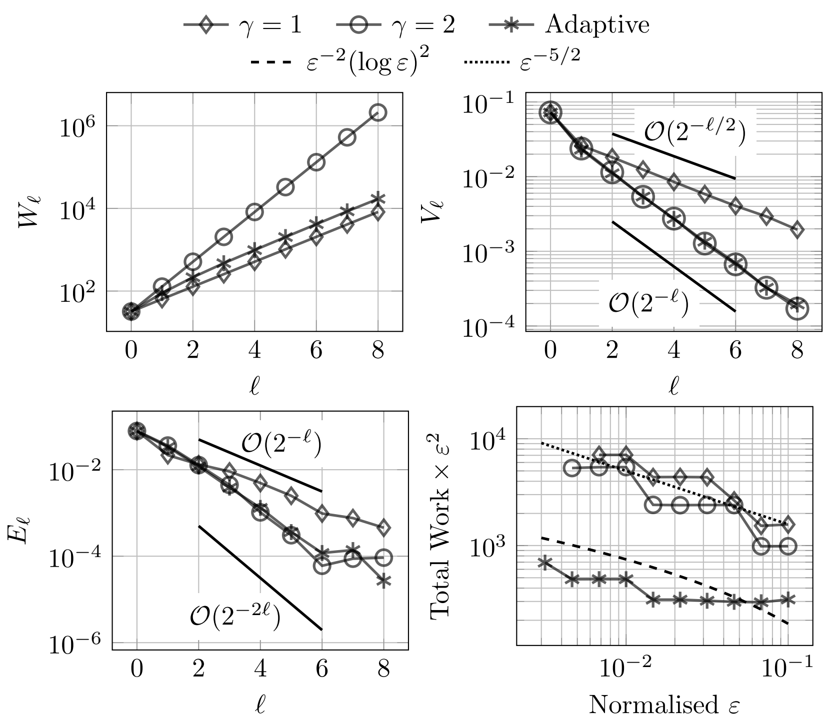

Fig. 4.1. *Results for the nested simulation model problem: Expected work per level W_ℓ (top left) taken as expected number of required inner samples from $X|Y$, multilevel correction variance V_ℓ (top right) and bias E_ℓ (bottom left) versus ℓ. The total work of MLMC times ε^2 versus ε, normalised by the true value of the solution (bottom right). Results are given for non-adaptive schemes with $\gamma = 1, 2$ and the adaptive scheme with $r = 1.95$.*

4.2. Stochastic Differential Equations. We now consider a setup where g is determined by d stock prices modeled by the geometric Brownian motions

$$\mathrm{d}S^{(i)}(t) = a_i S^{(i)}(t)\mathrm{d}t + b_i S^{(i)}(t)\mathrm{d}W^{(i)}(t), \quad 1 \le i \le d, \tag{4.3}$$

for constants $a_i, b_i \in \mathbb{R}$ and where the one-dimensional Brownian motions take the form

$$W^{(i)}(t) = \rho W_{\mathrm{com}}(t) + \sqrt{1-\rho^2}\, W^{(i)}_{\mathrm{ind}}(t),$$

for a correlation coefficient $\rho \in [0, 1]$ and independent Brownian motions $W_{\mathrm{com}}(t)$ and $\{W^{(i)}_{\mathrm{ind}}(t)\}_{i=1}^d$. Here, $W_{\mathrm{com}}(t)$ models common market noise shared by all of the stocks whereas $W^{(i)}_{\mathrm{ind}}(t)$ represents idiosyncratic noise of stock i only. Specifically, we set

$$g = \frac{1}{d}\sum_{i=1}^{d} S^{(i)}(1) - K, \tag{4.4}$$

so that $\mathbb{P}[\,g > 0\,]$ reflects the non-discounted price of a so-called digital option, a financial derivative which pays a unit price at time 1 if the mean value of the stocks exceeds K, and nothing otherwise.

The approximate samples, g_ℓ, are computed using either Euler-Maruyama or Milstein discretisation of the underlying SDEs with step size $h_\ell = 2^{-\gamma\ell}$. When adaptively refining samples of g_ℓ we use the Brownian Bridge construction to refine the sampled Brownian paths conditioned on their existing points [21, Section 1.8]. Specifically, given W_{nh_ℓ} and $W_{(n+1)h_\ell}$ we can sample the Brownian motion at time $(n+1/2)h_\ell$ using

$$W_{(n+1/2)h_\ell} \overset{\mathrm{d}}{=} \frac{W_{nh_\ell} + W_{(n+1)h_\ell}}{2} + \sqrt{\frac{h_\ell}{4}}\zeta_n, \quad \text{for } \zeta_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).$$

This procedure can be used recursively within Algorithm 3.1 to refine from step-size $h_{\ell+k}$ to $h_{\ell+k+1} = 2^{-\gamma} h_{\ell+k}$. It follows from the strong convergence results of each method that Assumption 1.3 holds for all $q < \infty$ using deterministic, constant σ_ℓ and $\beta = \gamma$ for Euler-Maruyama [21, Theorem 10.2.2] and $\beta = 2\gamma$ for the Milstein scheme [21, Theorem 10.3.5]. That Assumption 1.4 holds for constant $\sigma_\ell = \sigma$ can be shown for Euler-Maruyama using [17, Theorem 2.3] to bound the difference in the densities of g_ℓ and g. The result then follows since g has a bounded density [22, Theorem 10.9.11]. Moreover, from the weak convergence results in [21], we know that Assumption 2.7 holds for both SDE schemes with $\alpha = \beta$. Bounding the variance of $S^{(i)}(1)$ for all instances of $a_i, b_i, S^{(i)}(0)$ we see that $\mathrm{Var}[\,g\,] \propto d^{-1}$. Consequently, we choose $\sigma_\ell = \sigma = d^{-1/2}$.

We first consider (4.4) for a single stock, $d = 1$. In (4.3), we take $a_1 = 0.05, b_1 = 0.4$ and K is chosen such that $\mathbb{E}[\,\mathbb{H}(g)\,] = 0.025$. The terms W_ℓ, V_ℓ, C_ℓ of the MLMC estimator is shown in Figure 4.2 for Euler and Milstein approximation of g, using non-adaptive and adaptive simulation. For non-adaptive sampling we consider the cases $h_\ell = 2^{-\gamma\ell}$ for $\gamma = 1, 2$. The adaptive samplers take $\gamma = \theta = c = 1$. For the Euler-Maruyama scheme we take $r = 1.95$. Since $\beta = 2\gamma$ for the Milstein scheme, Lemma 3.4 allows us to take larger values of r and we set $r = 10$ here. W_ℓ is taken as the expected number of SDE steps required from the fine and coarse estimator at level ℓ. By construction, the work for both non-adaptive samplers is proportional to $2^{\gamma\ell}$. The adaptive schemes have $W_\ell \lesssim 2^\ell$ following Proposition 3.3. Note that the expected work per sample is slightly lower at each level for adaptive sampling using the Milstein scheme, as the larger value $r = 10$ requires fewer refinements to be made. For the Euler-Maruyama samplers we see $V_\ell \lesssim 2^{-\gamma\ell/2}$ for the non-adaptive and $V_\ell \lesssim 2^{-\ell}$ for the adaptive sampler, as expected for $\beta = \gamma$. These bounds are all squared when using the Milstein scheme since $\beta = 2\gamma$ in this case. Moreover, we observe $E_\ell \lesssim 2^{-\gamma\ell}$ for the non-adaptive sampler for both SDE schemes, with $E_\ell \lesssim 2^{-2\ell}$ for the adaptive samplers. This provides evidence that the stronger results following from Assumptions 2.6, 2.7 and 3.8 hold for the Euler-Maruyama scheme. For the Milstein scheme, the observed rates of E_ℓ follow immediately from the equivalent rates on V_ℓ and Remark 2.4.

For the non-adaptive schemes with $\gamma = 2$ and the adaptive samplers, we compute $\mathcal{M}^*$ for various error tolerances ε. In Figure 4.3, we plot the total work (taken as overall number of SDE time-steps) times ε^2 versus ε, normalized by the true solution. For the non-adaptive, Euler-Maruyama sampler, we observe a rate close to $\varepsilon^{-5/2}$ as predicted by Corollary 2.10. This is reduced to $\varepsilon^{-2}(\log\varepsilon)^2$ using adaptive sampling with the Euler-Maruyama scheme as in Theorem 3.10 for $q \to \infty$. Note that we

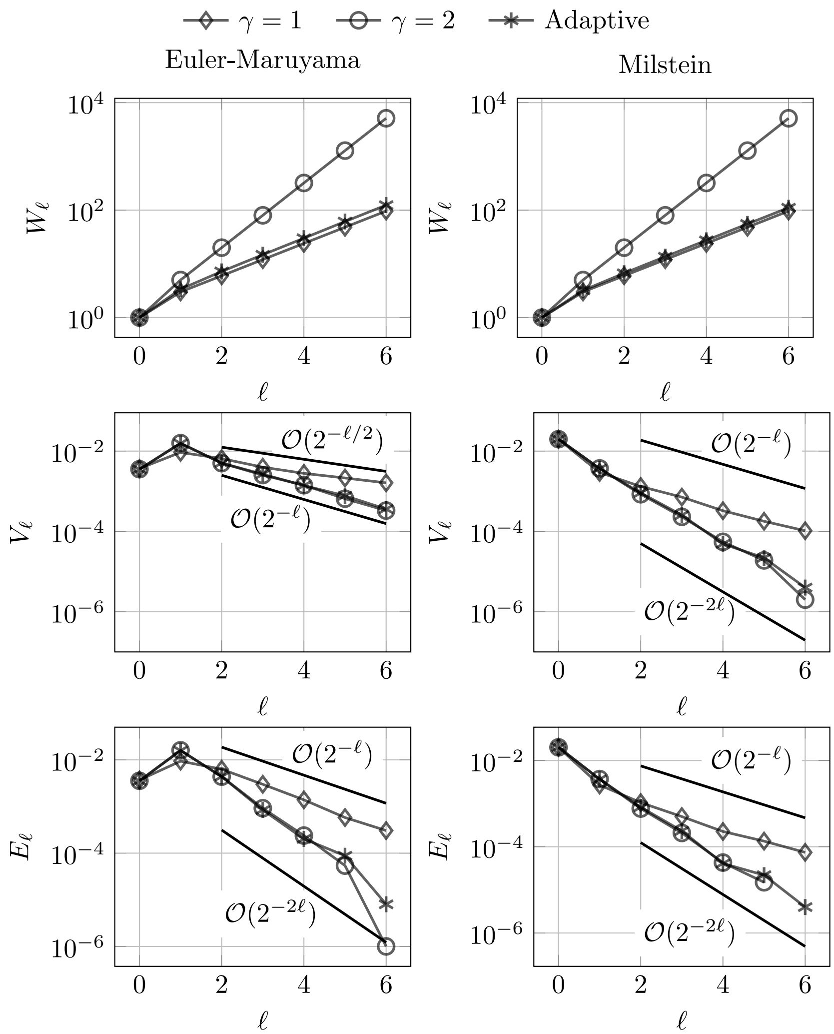

FIG. 4.2. *W_ℓ (top), V_ℓ (middle) and E_ℓ (bottom) vs ℓ for the one dimensional SDE problem using Euler-Maruyama (left) and Milstein (right) simulation of the underlying SDE. We consider non-adaptive samplers with $\gamma = 1, 2$ and adaptive sampling with $r = 1.95$ for the Euler scheme and $r = 10$ for Milstein simulation.*

observe the same rate without adaptive sampling when using the Milstein scheme, by Corollary 2.10 since $\beta = 2\gamma$. The cost is slightly lower for the non-adaptive Milstein scheme than for adaptive Euler-Maruyama, since the variance rate $V_\ell \lesssim 2^{-\gamma\ell}$ is observed without refining the samples beyond level ℓ at all. However, we obtain the best results by combining the Milstein scheme with adaptive MLMC. In this case we observe complexity very close to ε^{-2} as in Theorem 3.10.

To illustrate how this performance translates to higher dimensional problems we consider the case $d = 10$, with correlation coefficient $\rho = 0.2$. We assume that $0.05 \leq a_i \leq 0.15$, $0.01 \leq b_i \leq 0.4$ and $0.9 \leq S^{(i)}(0) \leq 1.1$, where the drift and diffusion coefficients a_i and b_i are uniformly sampled along with the initial values

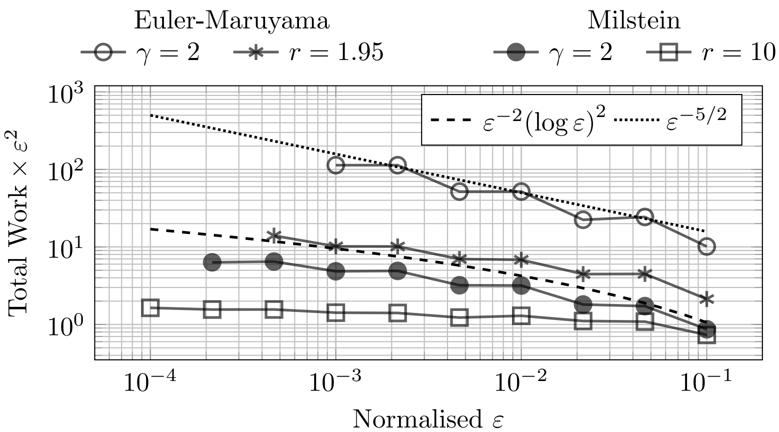

Fig. 4.3. *The total work times ε^2 of MLMC for the one-dimensional SDE problem versus ε, normalised by the true value of the solution. Results are shown for both Euler-Maruyama and Milstein simulation of the underlying SDE. For each method we show results for non-adaptive sampling with $\gamma = 2$ and adaptive sampling with $r = 1.95$ for the Euler scheme and $r = 10$ for Milstein simulation.*

$S^{(i)}(0)$. Euler-Maruyama simulation of g is used in non-adaptive MLMC for $\gamma = 1, 2$ and adaptive sampling with $\gamma = 1$ and $r = 1.95$. The parameter K is again tuned so that $\mathbb{E}[\,\mathbb{H}(g)\,] \approx 0.025$. The values of W_ℓ, V_ℓ, E_ℓ for each method are plotted against ℓ in Figure 4.4. We observe a slight increase to each term. However, the rates of each parameter are all equivalent to those seen before, and the complexity of MLMC is unaffected by the increased dimensionality. To emphasize this point, Figure 4.4 also displays the total work ($\times \varepsilon^2$) against ε. In particular, we again observe $\varepsilon^{-2}(\log \varepsilon)^2$ complexity for the adaptive sampler, as opposed to $\varepsilon^{-5/2}$ for the non-adaptive samplers.

5. Conclusion.

We presented an efficient, general, MLMC framework for computing probabilities as in (1.2). The inherent discontinuity in the problem leads to high complexities for standard MLMC methods. We are able to improve the performance of MLMC using adaptive sampling based on the methods for nested simulation in [13]. The approach used is applicable to a wide class of problems and is often able to recover the canonical ε^{-2} MLMC complexity. The theory is supported by numerical experiments for nested simulation and SDEs.

The adaptive algorithm is limited by a high kurtosis of the multilevel correction terms $\Delta\mathbb{H}_\ell$ caused by the discontinuous observable. This makes estimates of E_ℓ and V_ℓ unreliable for a small number of samples at larger levels. Smoothing methods [2,3,15] can control the kurtosis by removing the discontinuity. For the present work, Bayesian estimation [7,9] of E_ℓ and V_ℓ is used to improve the robustness of MLMC. Since a high kurtosis can have a large impact on the robustness of MLMC, the next step is to explore additional methods to reduce the kurtosis or obtain reliable estimates for E_ℓ and V_ℓ in this setup.

It is straightforward to extend the methods considered here to compute expectations of discontinuous functionals other than $\mathbb{I}_{G\in\Omega}$ or $\mathbb{H}(g)$. For example, in barrier option pricing, the payoff can be written as a product of a smooth/Lipschitz function with an indicator function. We will consider applications to other financial derivatives and risk measures in future work.

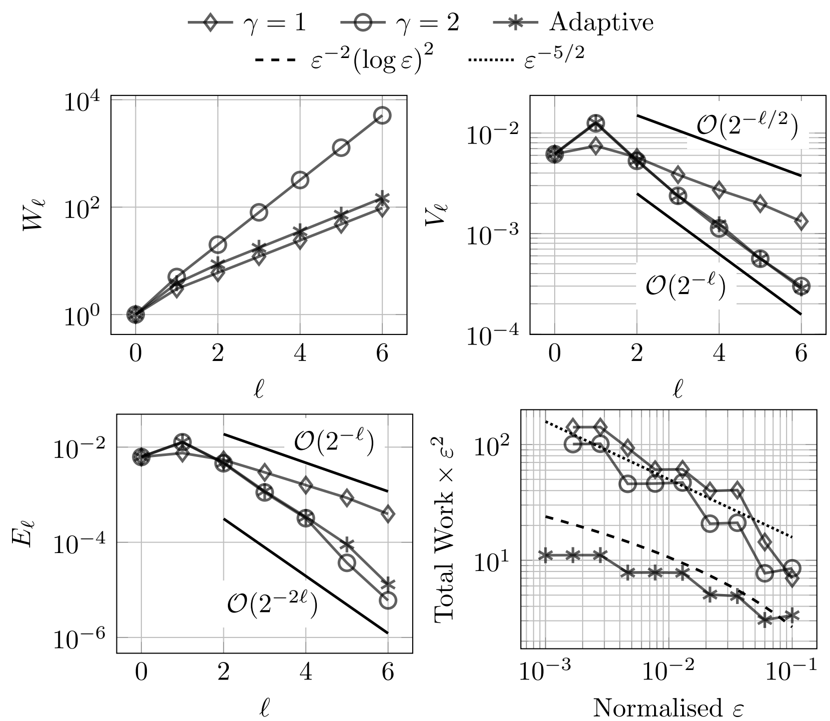

FIG. 4.4. *Results for 10-dimensional digital option: Expected work per level W_ℓ (top left) taken as expected number of time-steps required to sample g_ℓ and $g_{\ell-1}$, V_ℓ (top right) and E_ℓ (bottom left) versus ℓ. The total work of MLMC times ε^2 against ε, normalised by the true value of the solution is shown (bottom right). Results are given for non-adaptive schemes with $\gamma = 1, 2$ and the adaptive scheme with $r = 1.95$.*

Acknowledgments. We wish to acknowledge the helpful input and feedback received from Michael B. Giles throughout the development of this paper.

A-L. Haji-Ali was supported by a Sabbatical Grant from the Royal Society of Edinburgh.

J. Spence was supported by EPSRC grant EP/S023291/1.

REFERENCES

[1] R. AVIKAINEN, *On irregular functionals of SDEs and the Euler scheme*, Finance and Stochastics, 13 (2009), pp. 381–401, https://doi.org/10.1007/s00780-009-0099-7.

[2] C. BAYER, C. B. HAMMOUDA, AND R. TEMPONE, *Numerical smoothing and hierarchical approximations for efficient option pricing and density estimation*, 2020, https://arxiv.org/abs/2003.05708.

[3] F. BOURGEY, E. GOBET, AND C. REY, *A comparative study of polynomial-type chaos expansions for indicator functions*, (2021), https://hal.archives-ouvertes.fr/hal-03199734. preprint.

[4] F. BOURGEY, S. D. MARCO, E. GOBET, AND A. ZHOU, *Multilevel Monte Carlo methods and lower–upper bounds in initial margin computations*, Monte Carlo Methods and Applications, 26 (2020), pp. 131–161, https://doi.org/doi:10.1515/mcma-2020-2062, https://doi.org/10.1515/mcma-2020-2062.

[5] M. BROADIE, Y. DU, AND C. C. MOALLEMI, *Efficient risk estimation via nested sequential simulation*, Management Science, 57 (2011), pp. 1172–1194, https://doi.org/10.1287/mnsc.1110.1330.

[6] K. CLIFFE, M. B. GILES, R. SCHEICHL, AND A. TECKENTRUP, *Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients*, Computing and Visualization

in Science, 14 (2011), pp. 3–15, https://doi.org/10.1007/s00791-011-0160-x.

[7] N. Collier, A.-L. Haji-Ali, F. Nobile, E. von Schwerin, and R. Tempone, *A continuation Multilevel Monte Carlo algorithm*, BIT Numerical Mathematics, 55 (2015), pp. 399–432, https://doi.org/10.1007/s10543-014-0511-3.

[8] T. J. Dodwell, S. Kynaston, R. Butler, R. Haftka, N. H. Kim, and R. Scheichl, *Multilevel Monte Carlo simulations of composite structures with uncertain manufacturing defects*, Probabilistic Engineering Mechanics, 63 (2021), pp. 103–116, https://doi.org/10.1016/j.probengmech.2020.103116.

[9] D. Elfverson, F. Hellman, and A. Målqvist, *A Multilevel Monte Carlo Method for Computing Failure Probabilities*, SIAM/ASA Journal on Uncertainty Quantification, 4 (2016), pp. 312–330, https://doi.org/10.1137/140984294.

[10] M. B. Giles, *Improved Multilevel Monte Carlo Convergence using the Milstein Scheme*, in Monte Carlo and Quasi-Monte Carlo Methods 2006, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 343–358, https://doi.org/10.1007/978-3-540-74496-2_20.

[11] M. B. Giles, *Multilevel Monte Carlo Path Simulation*, Operations research, 56 (2008), pp. 607–617, https://doi.org/10.1287/opre.1070.0496.

[12] M. B. Giles, *Multilevel Monte Carlo methods*, Acta Numerica, 24 (2015), pp. 259—328, https://doi.org/10.1017/S096249291500001X.

[13] M. B. Giles and A.-L. Haji-Ali, *Multilevel nested simulation for efficient risk estimation*, SIAM/ASA Journal on Uncertainty Quantification, 7 (2019), pp. 497–525, https://doi.org/10.1137/18M1173186.

[14] M. B. Giles and A.-L. Haji-Ali, *Sub-sampling and other considerations for efficient risk estimation in large portfolios*, (2019), https://arxiv.org/abs/1912.05484. to appear in The Journal of Computational Finance.

[15] M. B. Giles, T. Nagapetyan, and K. Ritter, *Multilevel Monte Carlo approximation of distribution functions and densities*, SIAM/ASA Journal on Uncertainty Quantification, 3 (2015), pp. 267–295, https://doi.org/10.1137/140960086.

[16] D. Giorgi, V. Lemaire, and G. Pages, *Weak error for nested multilevel Monte Carlo*, Methodology and computing in applied probability, 22 (2020), pp. 1325–1348.

[17] E. Gobet and C. Labart, *Sharp estimates for the convergence of the density of the Euler scheme in small time*, Electronic Communications in Probability, 13 (2008), pp. 352–363, https://doi.org/10.1214/ECP.v13-1393.

[18] M. B. Gordy and S. Juneja, *Nested Simulation in Portfolio Risk Measurement*, Management Science, 56 (2010), pp. 1833–1848, https://doi.org/10.1287/mnsc.1100.1213.

[19] J. Gregory, *Counterparty credit risk and credit value adjustment: a continuing challenge for global financial markets*, Wiley Finance Series, Wiley, 2nd edition. ed., 2012.

[20] F. Jonsson, *On the heavy-tailedness of Student's t-statistic*, Bernoulli, 17 (2011), pp. 276–289.

[21] P. E. Kloeden, *Numerical solution of stochastic differential equations*, Applications of mathematics; 23, Springer, Berlin ; New York, corrected third printing ed., 1999.

[22] H.-H. Kuo, *Introduction to Stochastic Integration*, Universitext, Springer, 2006.

[23] S. Taverniers, S. B. M. Bosma, and D. M. Tartakovsky, *Accelerated multilevel Monte Carlo with kernel-based smoothing and Latinized stratification*, Water resources research, 56 (2020).

[24] Z. Xu, Z. He, and X. Wang, *Efficient risk estimation via nested multilevel quasi-Monte Carlo simulation*, 2020, https://arxiv.org/abs/2011.11898.

Appendix A. Discussion of Assumption 3.8.

This appendix discusses the necessity of Assumption 3.8 to observe better convergence rates for E_ℓ due to adaptive sampling. Evidence is given in the form of a numerical experiment when Assumption 3.8 is false. In particular, we consider (1.2) where $g \sim \mathcal{N}(-\mu, 1)$ for $\mu > 0$ chosen such that $\mathbb{P}[\, g > 0\,] = 0.025$. Approximations g_ℓ are artificially sampled through

$$g_\ell = g + 2^{-\ell\gamma/2}\Big(2^{-\ell\gamma/2} + \zeta^2 - 1\Big), \quad \zeta \sim \mathcal{N}(0, 1).$$

We assume an artificial cost of $2^{\gamma\ell}$ in sampling g_ℓ. It follows that Assumptions 1.3 and 1.4 hold for $\beta = \gamma$, $\sigma_\ell \equiv \sigma$ constant and any $q < \infty$. From (1.4),

$$Z_\ell = \frac{2^{-\ell\gamma/2} + \zeta^2 - 1}{\sigma},$$

and so Assumption 2.7 holds for $\beta = \gamma$. However, Assumption 3.8 is false since

$\mathbb{P}[\, Z_\ell > 0\,] \to 1$ as $\ell \to \infty$. Thus, the hypothesis of Lemma 3.9 is false.

Figure A.1 plots E_ℓ as in (2.2) for non-adaptive sampling with $\gamma = 1, 2$ and for adaptive sampling with $r = 1.95, \gamma = \theta = c = 1$ and $\sigma_\ell = \sigma = \sqrt{3}$. We use the same sample of ζ for the fine and coarse levels in each MLMC correction term, and when adaptively refining samples. For all methods we see $E_\ell \lesssim 2^{-\gamma\ell}$. Since $\beta = \gamma$ this agrees with Proposition 2.9 for the non-adaptive samplers. If Assumption 3.8 holds, we expect that the adaptive refinement would provide $E_\ell \lesssim 2^{-2\ell}$ by Lemma 3.9. However, we observe a worse rate numerically, which provides evidence that Assumption 3.8 is crucial to improve the bias convergence rate using adaptive sampling.

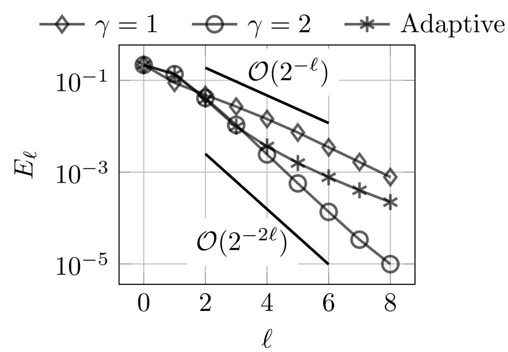

FIG. A.1. *E_ℓ versus ℓ for the artificial problem considered in Appendix* A *used as evidence of worse weak error rates for adaptive sampling when Assumption* 3.8 *is false.*

Appendix B. Different values of σ_ℓ. In Subsection 4.1 we assumed σ_ℓ^2 was the sample conditional variance of X given Y. However, the other examples considered all use constant values $\sigma_\ell \equiv \sigma$. In this appendix we discuss other choices of σ_ℓ for the nested simulation problem and the impact on the work of MLMC. One option is to take $\sigma_\ell^2 = \mathrm{Var}[\, X|Y\,]$. However, this information is likely unavailable for all practical applications. Thus, we consider instead the approximation $\sigma_\ell \equiv \sigma$, for some constant $\sigma > 0$. For Assumption 1.3 to hold for $\sigma_\ell \equiv \sigma$, we now require bounded moments of $g - g_\ell$, opposed to the t-statistic appearing in (4.2), which is a more restrictive condition.

We present results using adaptive sampling for the model problem presented in Subsection 4.1. Specifically, we estimate the total work of MLMC with several error tolerances ε and optimal starting levels as in Subsection 4.1 but with constant $\sigma_\ell = \sigma$. The total work required with constant $\sigma_\ell = \sigma$ divided by the work when using the sample standard deviation is shown in Figure B.1. The solid markers show the value σ^2 estimating the value $\mathbb{E}[\,\mathrm{Var}[\, X|Y\,]\,]$. When $\sigma \to 0$, the term $|\delta_\ell|$ in Algorithm 3.1 tends to ∞ and we instead use deterministic sampling with $N_\ell = N_0 2^\ell$ inner samples per level in the limiting case. Conversely, when $\sigma \to \infty$, $|\delta_\ell| \to 0$ and the adaptive algorithm reverts to deterministic sampling with $N_\ell = N_0 2^{2\ell}$ inner samples per level. This leads to expensive pre-asymptotic regimes for large and small σ and we observe worse performance as ε decreases. The work is typically lower for large σ opposed to small σ is consistent with results showing MLMC is more effective when the approximations are refined by a factor of around 7 per level in this application ($W_\ell \propto 7^\ell$) [11]. The only value of σ for which we consistently observe equal performance using constant σ_ℓ, opposed to the sample variance, is $\sigma^2 \approx \mathbb{E}[\,\mathrm{Var}[\,X|Y\,]\,]$. MLMC actually has slightly lower cost for constant σ_ℓ in this instance, likely due to statistical errors in the sample variance impacting the refinement of certain samples, whereas taking constant $\sigma^2 \approx \mathbb{E}[\,\mathrm{Var}[\,X|Y\,]\,]$ refines samples enough on average to observe the benefits of adaptive sampling. To draw further conclusions, more rigorous justification of Assumptions 1.3 and 1.4 is required.

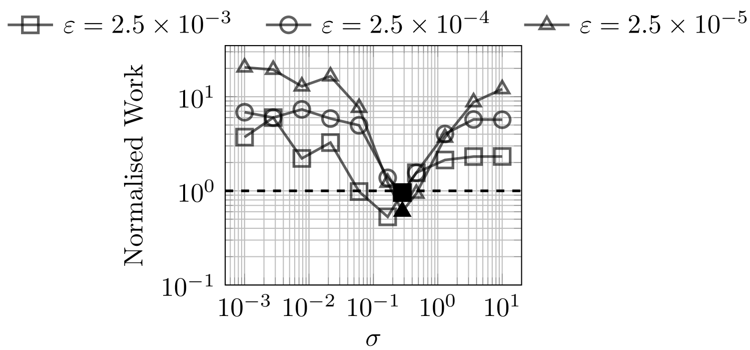

FIG. B.1. *The work required for the model problem in Subsection* 4.1 *using adaptive MLMC with constant $\sigma_\ell = \sigma$, normalised by the work when σ_ℓ is the sample standard deviation. The solid markers show $\sigma = \sqrt{\mathbb{E}[\,\mathrm{Var}[\,X|Y\,]\,]} \approx 0.28$.*