Strong Antithetic Variates: Theory and Applications

Abolfazl Hashemi, Dongmin Lee, and Anuran Makur

***Abstract*—The antithetic variates method is a well-known variance reduction technique for Monte Carlo sampling that is known to be empirically effective in many settings. However, precise theoretical analyses that quantify the degree of variance reduction remain unexplored. In this work, as a step towards developing such theoretical guarantees, we show how strongly isotonic assumptions allow us to derive stronger antithetic variance reduction inequalities that provide quantitative lower bounds on the degree of variance reduction. To this end, we develop the closely intertwined concepts of antithetic maps and indices, and reveal their useful properties and ties to problems in optimal transport and signal processing. In addition, we use our stronger antithetic variance reduction inequalities to demonstrate improved theoretical guarantees when antithetic variates are used in various applications. These applications encompass various topics including approximating integrals, function approximation, concentration inequalities, and stochastic optimization. Our arguments utilize and develop ideas from correlation inequalities and first-order optimization theory among other tools.**

***Index Terms*—Variance reduction, antithetic variates, function approximation, stochastic optimization.**

I. Introduction

Variance reduction is a major topic of theoretical and practical interest in the field of Monte Carlo sampling [2]–[7], which is widely used in signal processing and machine learning [8]–[12]. One simple yet effective method of variance reduction is to use *antithetic variates* [13]. The idea behind antithetic variates (and some other variance reduction techniques such as the control variates method [14]) is to introduce auxiliary variables that are negatively correlated to the original samples so that their addition "cancels out" the variance.

For a simple example, suppose the objective is to estimate $\int_0^1 g(t)\,\mathrm{d}t$. The standard Monte Carlo method would be to generate n independent and identically distributed (i.i.d.) uniform random samples $U_1, \ldots, U_n \sim \mathsf{Unif}(0,1)$ and evaluate the canonical estimator $\frac{1}{n}\sum_{i=1}^n g(U_i)$. Antithetic variates improve on this by sampling g at $1-U_i$ in addition to U_i, hoping to induce negative correlation in the summands while keeping the estimator unbiased [15], [16] (also see [7, Chapter 5.2]), i.e., $\frac{1}{2n}\sum_{i=1}^n \left(g(U_i)+g(1-U_i)\right)$.

Since $\operatorname{cov}(g(U_i), g(1-U_i)) \leq \operatorname{var}(g(U_i))$, where $\operatorname{var}(\cdot)$ and $\operatorname{cov}(\cdot,\cdot)$ denote the variance and covariance operators, respectively, it is trivial to show that the antithetic estimator has no greater variance than the canonical Monte Carlo estimator with n i.i.d. samples. Thus, antithetic variates achieve a variance reduction while using the same amount of randomness (number of independent random samples).

However, oftentimes the metric of concern is the number of function evaluations (e.g., in oracle complexity analyses), in which case it is more appropriate to compare the antithetic estimator with a canonical estimator that uses $2n$ i.i.d. samples. In this regime, it is well-known that antithetic variates achieve variance reduction when $g:[0,1]\to\mathbb{R}$ is monotonic [7]:

$$\operatorname{var}\left(\frac{1}{2n}\sum_{i=1}^{n}\left(g(U_i)+g(1-U_i)\right)\right) \leq \operatorname{var}\left(\frac{1}{2n}\sum_{i=1}^{2n} g(U_i)\right). \quad (1)$$

Furthermore, the antithetic variates method can also be adapted to general distributions via the antithetic estimator (cf. [7]):

$$\frac{1}{2n}\sum_{i=1}^{n}\left(g(Z_i)+g(F^{-1}(1-F(Z_i)))\right), \quad (2)$$

where F is the cumulative distribution function (CDF) of the continuous random variable $Z\in\mathbb{R}$ and F^{-1} is its generalized inverse (see Definition 1).

Note that (1) does not provide any guarantees on the magnitude of the variance reduction that can be achieved with antithetic variates. Such guarantees are of significant theoretical interest, as they can be used to improve the theoretical bounds of many stochastic results and algorithms that are sensitive to variance, e.g., Monte Carlo integration [3], concentration inequalities [17], stochastic optimization [18], [19], and function approximation [20], [21]. However, such theoretical guarantees are often not available in the literature, because merely assuming that g is monotonic is too weak to always ensure a meaningful reduction in variance. Indeed, in the extreme case when g is constant (which is technically monotonic), there is no variance reduction (nor any variance to be reduced in the first place).

Motivated by this observation, we use the *anti-Lipschitzness* of a univariate function or *strong isotonicity* of a multivariate function, as defined in Assumptions 3 and 4, to capture the "strength" of its monotonicity. In addition, to derive general inequalities that can be easily applied to a wide variety of probability distributions, we define a measure of self-anticorrelation between a random variable and its antithetic counterpart, the *antithetic index*, to isolate the sample distribution's contribution to the variance reduction.

A. Main Contributions

We next delineate our main contributions.

1) We define the *antithetic map* (Definition 2), which maps a random variable to its antithetic counterpart, and the

The author ordering is alphabetical. This work was supported in part by the National Science Foundation (NSF) under Grant CNS-2313109 and in part by the NSF CAREER Award under Grant CCF-2337808. This work was presented in part at the 2025 IEEE International Symposium on Information Theory (ISIT) [1].

Abolfazl Hashemi is with the Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA (e-mail: abolfazl@purdue.edu).

Dongmin Lee is with the Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA (e-mail: lee4818@purdue.edu).

Anuran Makur is with the Department of Computer Science and the Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA (e-mail: amakur@purdue.edu).

antithetic index (Definition 3), an intrinsic property of probability distributions that measures the magnitude of negative covariance between an antithetic pair.

2) We analyze some properties of the antithetic map and index (Theorem 2) and evaluate values for some common probability distributions (Table I). Furthermore, we present some observations that reveal interesting links between antithetic variates and various other problems in optimal transport and signal processing (Theorem 3).
3) We find a lower bound on the effectiveness of the antithetic variates method that depends on the anti-Lipschitz constant (see Assumption 3) and the antithetic index of the random variable (Theorem 4).
4) We generalize our results to multivariate functions and establish a *strong antithetic variance reduction inequality* (Theorems 5 and 6).
5) Finally, we investigate several applications of sampling methods and show that their theoretical guarantees can be improved with antithetic variates:
   a) Monte Carlo integration (Proposition 7),
   b) Function approximation (Proposition 8),
   c) Concentration inequalities such as Bernstein's inequality (Propositions 9 and 11), Bennett's inequality (Proposition 12), and Lipschitz concentration for Gaussian random variables (Proposition 13),
   d) Stochastic optimization in the nonconvex (Theorem 14) and strongly convex (Theorem 15) settings.

B. Related Literature

Here we provide a brief summary of the existing literature on antithetic variates, variance reduction, and some of the applications that we examine in Sections II-D to II-F. The concept of antithetic variates was introduced in [13] in a line of work that focused on harnessing correlation to reduce variance (see [2] for an early survey). Further research by [15], [16], [22]–[24] analyzed other theoretical properties of antithetic variates. Though some of the earlier Monte Carlo literature did not mention the resemblance, the manner in which antithetic methods harnessed negative correlation resembled known correlation inequalities (see, e.g., [25]–[27]). We develop and exploit such ideas in our proofs. More recently, [19], [28] have studied how antithetic methods can help empirically improve the performance of sampling algorithms in machine learning. We refer readers to textbooks such as [5, Chapter V] or [7, Chapter 5] for a broader overview of variance reduction methods.

One of the applications of antithetic variates that we examine in Section II-F is variance-reduced stochastic optimization, which has attracted significant attention since the advent of methods such as Stochastic Average Gradient (SAG) [29], [30], SAGA [31], Stochastic Variance Reduced Gradient (SVRG) [18], [32], [33], nonparametric regression methods [34], and non-uniform (importance) sampling [35]–[37] (see [38] for a comprehensive survey). Another problem that we analyze in Section II-D is function approximation, which has a rich literature with some machine-learning-related results tracing its roots back to [20], [21], [39]–[41].

C. Outline

First, in Section II-A, we begin with an overview of the notation and conventions used in this paper and introduce concepts such as the generalized inverse and strongly isotonic functions. Next, in Section II-B, we define the antithetic map and index and examine their properties. In Section II-C, we present our main results on strong antithetic variance reduction inequalities. In Sections II-D to II-F, we use our strong variance reduction inequalities to derive strengthened theoretical guarantees for various applications including function approximation, concentration of measure, and stochastic gradient descent. In Sections III to VII, we present the proofs of the aforementioned results. Finally, in Section VIII, we conclude by summarizing our results and discuss future research directions.

II. Main Results

A. Preliminaries

We begin by introducing the conventions and assumptions used throughout this paper. First, we impose some basic constraints on the random variables we study in order to avoid pathological cases.

Assumption 1 (Non-atomic distribution)**.** The random variable $Z \in \mathbb{R}$ is non-atomic, i.e., it has continuous CDF $F : \mathbb{R} \to [0,1]$.

Assumption 2 (Finite second moment)**.** The random variable $Z \in \mathbb{R}$ has finite second moment, i.e., $\mathbb{E}\left[Z^2\right] < +\infty$.

Although we do not mention it explicitly, Borel measurability is assumed implicitly for various functions and random variables in the paper. Since F is not necessarily strictly increasing, its inverse might not be well defined. Thus, we use the notion of a generalized inverse (or *quantile function*); see [42] for an exposition of this widely adopted definition.

Definition 1 (Generalized inverse [42])**.** The *generalized inverse* $F^{-1} : [0,1] \to \overline{\mathbb{R}}$ of the CDF $F : \mathbb{R} \to [0,1]$ of a real-valued random variable Z is defined as $F^{-1}(y) \triangleq \inf\{x \in \mathbb{R} : F(x) \geq y\}$, where we define $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ and let $\inf \varnothing = +\infty$ and $\inf \mathbb{R} = -\infty$.

Note that if Assumption 1 is satisfied, there must exist an x such that $F(x) = y$ for any $y \in (0,1)$. Thus, if Z is non-atomic, $F(F^{-1}(y)) = y$ for all $y \in [0,1]$ when we let $F(-\infty) = 0$ and $F(+\infty) = 1$. Note also that F^{-1} agrees with the usual definition of inverse on the range of F when F is strictly increasing and continuous [42], and that $F^{-1}(F(Z)) = Z$ a.s. (almost surely).

A main theme of this work is to exploit strong monotonicity assumptions on a function to prove stronger theoretical guarantees. Strong monotonicity can be formally defined in the univariate and multivariate settings as follows; the definitions are presented as assumptions for convenience in the sequel.

Assumption 3 (Monotone increasing anti-Lipschitz function)**.** The function $g : \mathbb{R} \to \mathbb{R}$ is monotone increasing and *c-anti-Lipschitz* (see [43, Definition 4.4]) with constant $c > 0$ if for every $x > y$, $g(x) - g(y) \geq c(x - y)$.

Assumption 4 (Strongly isotonic function)**.** The function $g : \mathbb{R}^d \to \mathbb{R}$ is *c-strongly isotonic* with constant $c > 0$ if for

all $\boldsymbol{x} = (x_1, \ldots, x_d), \boldsymbol{y} = (y_1, \ldots, y_d) \in \mathbb{R}^d$ with $\boldsymbol{x} \geq \boldsymbol{y}$ (i.e., $x_i \geq y_i$ for all $i \in \{1, \ldots, d\}$), we have $g(\boldsymbol{x}) - g(\boldsymbol{y}) \geq c \|\boldsymbol{x} - \boldsymbol{y}\|_\infty$, where $\|\cdot\|_q$ denotes the ℓ^q-norm for $q \in [1, \infty]$. Furthermore, $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^p$, $\boldsymbol{g}(\boldsymbol{x}) = (g_1(\boldsymbol{x}), \ldots, g_p(\boldsymbol{x}))$ is *coordinate-wise c-strongly isotonic* if g_i is c-strongly isotonic for all $i \in \{1, \ldots, p\}$.

As the following proposition shows, such strongly isotonic functions are closely related to strongly monotone operators, where $f : \mathbb{R}^d \to \mathbb{R}^d$ is said to be c-strongly monotone (with $c > 0$) if $(f(\boldsymbol{x}) - f(\boldsymbol{y})^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{y}) \geq c \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ or simply monotone if $c = 0$ (see, e.g., [44]).

Proposition 1 (Isotonic functions and monotone operators)**.** *Given a function $g : \mathbb{R}^d \to \mathbb{R}$, if there exists an entry-wise positive vector $\boldsymbol{u} \in \mathbb{R}^d$ such that $g(\boldsymbol{x})\boldsymbol{u}$ is a (strongly) monotone operator, then g is (strongly) isotonic.*

Proposition 1 is proved in Appendix A. Note that a conservative monotone operator $g(\boldsymbol{x})\boldsymbol{u}$ can be perceived as the gradient of a convex ridge function $F : \mathbb{R}^d \to \mathbb{R}$ with the form $F(\boldsymbol{x}) = G(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x})$ for some scalar function $G : \mathbb{R} \to \mathbb{R}$ [45].

B. Antithetic Indices

In order to derive results for general probability distributions, we begin by revisiting the antithetic variates method (cf. (2)) and analyze the *antithetic map* $F^{-1}(1 - F(Z))$.

Definition 2 (Antithetic map)**.** The **antithetic map** $a_Z : \mathbb{R} \to \overline{\mathbb{R}}$ of a random variable $Z \in \mathbb{R}$ with CDF F is defined as

$$a_Z(z) \triangleq F^{-1}(1 - F(z)),$$

where F^{-1} is the generalized inverse of F as defined in Definition 1. The random variable $a_Z(Z)$ is referred to as the antithetic pair (or antithetic counterpart) of Z. Note that although the codomain of a_Z is the extended reals, this does not cause any issues because $a_Z(Z) \in \mathbb{R}$ a.s.

As we will soon show, a_Z preserves the distribution of Z while making the pair $(Z, a_Z(Z))$ as anticorrelated as possible. We introduce the concept of *antithetic indices* to quantify the degree of this inverse correlation.

Definition 3 (Antithetic index)**.** The **antithetic index** $\tau(Z)$ of a random variable $Z \in \mathbb{R}$ with CDF F and finite second moment is defined as

$$\begin{aligned}\tau(Z) &\triangleq -\mathrm{cov}(Z, a_Z(Z)) \\ &= \left(\int_0^1 F^{-1}(u)\,\mathrm{d}u\right)^2 - \int_0^1 F^{-1}(u)F^{-1}(1-u)\,\mathrm{d}u.\end{aligned}$$

For CDFs where the mean exists and is finite and the second moment is infinite, the antithetic index remains well-defined, but it can be infinite. Note that we omit Z and denote the antithetic map and index as a and τ whenever it is unambiguous. The next theorem presents several useful properties of $\tau(Z)$.

Theorem 2 (Properties of the antithetic index)**.** *The antithetic map a_Z and antithetic index $\tau(Z)$ of a random variable $Z \in \mathbb{R}$ with CDF $F : \mathbb{R} \to [0, 1]$ satisfy the following properties:*

(a) $a_Z(Z)$ and Z are identically distributed if the distribution of Z is non-atomic.

(b) a_Z is monotone non-increasing. Furthermore, if F is continuous and strictly increasing in the support of Z, a_Z is continuous and strictly decreasing in the support of Z.

(c) If Z has strictly increasing and continuous CDF, a_Z is the unique function that satisfies properties (a) and (b).

(d) For any $\alpha \neq 0$ and $\beta \in \mathbb{R}$, $Y = \alpha Z + \beta$ has antithetic map $a_Y(Y) = \alpha a_Z(\alpha^{-1}(Y - \beta)) + \beta$. If Z has finite second moment, $Y = \alpha Z + \beta$ has antithetic index $\tau(Y) = \alpha^2 \tau(Z)$, which implies that $\tau(Y)/\mathrm{var}(Y) = \tau(Z)/\mathrm{var}(Z)$.

(e) $\tau(Z) \geq 0$. Arbitrarily small τ for fixed variance σ^2 can be achieved by a piecewise uniform random variable Z_t parameterized by $t \in \left[\frac{1}{2}, 1\right)$ and supported on $\left[0, \sigma\sqrt{3/(t(1-t))}\right]$ with probability density function (PDF)

$$f_t(z) = \begin{cases} \frac{1}{\sigma}\frac{1-t}{t}\sqrt{\frac{t(1-t)}{3}}, & \text{if } 0 \leq \frac{1}{\sigma}\sqrt{\frac{t(1-t)}{3}}z \leq t, \\ \frac{1}{\sigma}\frac{t}{1-t}\sqrt{\frac{t(1-t)}{3}}, & \text{if } t \leq \frac{1}{\sigma}\sqrt{\frac{t(1-t)}{3}}z \leq 1, \end{cases}$$

which has variance $\mathrm{var}(Z_t) = \sigma^2$ and $\lim_{t \to 1} \tau(Z_t) = 0$.

(f) If Z is non-atomic and has finite second moment, $\tau(Z) \leq \mathrm{var}(Z)$, with equality if and only if F is symmetric, i.e., there exists some $d \in \mathbb{R}$ such that $F(d - z) = 1 - F(z)$ for all $z \in \mathbb{R}$.

(g) For any compact interval $I = [\alpha, \alpha + L]$ of length L,

$$\sup_{Z:\, \mathit{support}(Z) = I} \tau(Z) = \frac{L^2}{4},$$

where the supremum can be achieved by a sequence of L-scaled symmetric beta random variables with both shape parameters tending to 0, i.e., if $Z_k = \alpha + LB_k$ where $B_k \sim \mathsf{Beta}(1/k, 1/k)$, $\lim_{k \to \infty} \tau(Z_k) = L^2/4$.

Theorem 2 is proved in Section III-A. Note that although the antithetic map and index remain well defined even for distributions with atoms, desirable properties such as Theorem 2(a) (which is required for the antithetic estimator to remain unbiased) are no longer guaranteed, which is why we assume non-atomicity in most of our work.

The antithetic index is an intrinsic property of probability distributions that can be readily computed from the CDF and quantile function using either analytical or numerical integration. For example, when $Z \sim \mathsf{Unif}(0, 1)$, we obtain $\tau(Z) = -\mathrm{cov}(Z, a(Z)) = -\mathrm{cov}(Z, 1 - Z) = \mathbb{E}[Z^2] - \mathbb{E}[Z](1 - \mathbb{E}[Z]) = \mathrm{var}(Z) = \frac{1}{12}$. As another example, when $Z \sim \mathcal{N}(0, 1)$ is standard Gaussian, we obtain $\tau(Z) = -\mathrm{cov}(Z, a(Z)) = -\mathrm{cov}(Z, -Z) = \mathrm{var}(Z) = 1$. For these two examples, where the distribution is symmetric, the antithetic index is equal to the variance due to Theorem 2(f). On the other hand, for random variables with asymmetric probability distributions, such as exponential $Z \sim \mathsf{Exp}(1)$, the antithetic index is strictly less than the variance: $\tau(Z) = -\int_0^\infty (z - 1)(-\log(1 - e^{-z}) - 1)(e^{-z})\,\mathrm{d}z = \frac{\pi^2}{6} - 1 < 1 = \mathrm{var}(Z)$.

Table I summarizes the antithetic indices and variances of some common symmetric and asymmetric distributions including the previously discussed examples. Intuitively, the antithetic index can be perceived as measuring the "intrinsic negative correlation" of a probability distribution, which is

TABLE I
A LIST OF SOME PROBABILITY DISTRIBUTIONS ALONG WITH THEIR ANTITHETIC INDICES AND VARIANCES. NOTE THAT $\tau(Z) = \mathrm{var}(Z)$ FOR SYMMETRIC DISTRIBUTIONS, BUT NOT FOR ASYMMETRIC ONES.

Distribution	Antithetic Index	Variance	$\tau(Z)/\mathrm{var}(Z)$
$\mathsf{Unif}(a,b)$	$\frac{(b-a)^2}{12}$	$\frac{(b-a)^2}{12}$	1
$\mathcal{N}(\mu,\sigma)$	σ^2	σ^2	1
$\mathsf{Exp}(\lambda)$	$\frac{\pi^2-6}{6\lambda^2}$	$\frac{1}{\lambda^2}$	$\frac{\pi^2}{6} - 1 \approx 0.645$
$\mathsf{Beta}(\alpha,1)$	$\left(\frac{\alpha}{\alpha+1}\right)^2 - \frac{B(\frac{1}{\alpha},\frac{1}{\alpha})}{2(\alpha+2)}$	$\frac{\alpha}{(\alpha+1)^2(\alpha+2)}$	$\frac{2\alpha^2(\alpha+2)-(\alpha+1)^2B(\frac{1}{\alpha},\frac{1}{\alpha})}{2\alpha}$

maximized when the distribution is symmetric and minimized when it differs drastically from its reflection.

Furthermore, the antithetic map can alternatively be characterized as a maximum correlation coupling between a probability distribution and its mirror image (also see "maximal correlation" in statistics and information theory [46]). This formulation also reveals close ties between antithetic variates and optimal transport theory, as the following theorem details.

Theorem 3 (Maximal correlation and optimal transport formulations of antithetic variates)**.** *The antithetic map a_Z and antithetic index $\tau(Z)$ of a non-atomic random variable $Z \sim P$ with finite second moment has the following properties:*

(a) The antithetic map $-a_Z$ is the maximum correlation coupling *between the distributions of Z and $-Z$, i.e.,*

$$-a_Z = \operatorname*{arg\,max}_{f:\mathbb{R}\to\mathbb{R}, -f(Z)\sim P} \mathrm{corr}(Z, f(Z)).$$

(b) Let Q be the distribution of $2\mathbb{E}[Z] - Z$, i.e., the reflection of P with respect to its mean. Then,

$$W_2(P,Q) = \sqrt{2\mathrm{var}(Z) - 2\tau(Z)},$$

where $W_2(\cdot,\cdot)$ is the 2-Wasserstein distance [47]:

$$W_2(P,Q) = \inf_{\gamma} \sqrt{\mathbb{E}_{(X,Y)\sim\gamma}[(X-Y)^2]}$$

where the infimum is over all couplings (joint distributions) of $X \sim P$ and $Y \sim Q$ with P and Q as marginals.

(c) Let R be the distribution of $(Z + a_Z(Z))/2$. Then,

$$W_2(P,R) \le \sqrt{\frac{\mathrm{var}(Z) + \tau(Z)}{2}}.$$

(d) Let S be the distribution of $(Z - a_Z(Z))/2 + \mathbb{E}[Z]$. Then,

$$W_2(P,S) \le \sqrt{\frac{\mathrm{var}(Z) - \tau(Z)}{2}} = \frac{W_2(P,Q)}{2}.$$

Theorem 3 is proved in Section III-B. Notably, the optimal transport formulation of the antithetic map provides an explanation for why the non-atomicity assumption is needed in order for the antithetic map to be well-defined. Although the Kantorovich formulation of the optimal transport problem (which minimizes over transport plans instead of transport maps) is known to be equivalent to Monge's original formulation for non-atomic distributions [47], this is not true in general for distributions with atoms. Generalizing the antithetic map to an antithetic "plan" (or equivalently a stochastic map) might accommodate for such distributions, but this is outside the scope of this paper.

In addition, the antithetic map has an interesting connection to another classical signal processing problem. The proof of Theorem 3(b) reveals that $W_2(P,Q)^2 = 4\mathrm{var}\big(\frac{1}{2}(Z + a_Z(Z))\big)$, while Theorem 3(c) and Theorem 3(d) present other relations between Z, $\frac{1}{2}(Z + a_Z(Z))$, and $\frac{1}{2}(Z - a_Z(Z))$. In general, the transformation $\frac{1}{2}(Z + a_Z(Z))$, which is closely related to the antithetic index, may discard information about the distribution of Z. So, an interesting question is: *When can we recover the distribution of Z from that of $Y = \frac{1}{2}\big(Z + a_Z(Z)\big)$?* To simplify this question, let U be a uniform random variable on the interval $\big(-\frac{1}{2}, \frac{1}{2}\big)$. Since $Z + a_Z(Z) = Z + F^{-1}(1 - F(Z)) \overset{\mathrm{d}}{=} F^{-1}\big(\frac{1}{2} + U\big) + F^{-1}\big(\frac{1}{2} - U\big)$, where $\overset{\mathrm{d}}{=}$ denotes equality in distribution, define the monotone non-decreasing function $g : \big(-\frac{1}{2}, \frac{1}{2}\big) \to \mathbb{R}$, $g(x) = F^{-1}\big(x + \frac{1}{2}\big)$ and let $g_{\mathsf{e}}(x) = \frac{1}{2}(g(x) + g(-x))$ be the even part of g. Then, $Z \overset{\mathrm{d}}{=} g(U)$ and $Y \overset{\mathrm{d}}{=} g_{\mathsf{e}}(U)$. We seek to recover the distribution of $g(U)$ from that of $g_{\mathsf{e}}(U)$, which is challenging because g_{e} is not usually monotone (and Y's CDF is a monotone rearrangement of g_{e} that is difficult to analyze). So, we consider the simpler question: *When can we recover g from g_{e}?*

To answer this latter question, denote the Fourier transform of g as $G(\omega) = \int_{-1/2}^{1/2} g(x) e^{-i\omega x}\,\mathrm{d}x$ (for $\omega \in \mathbb{R}$), which is well-defined because

$$\int_{-1/2}^{1/2} |g(x)e^{-i\omega x}|\,\mathrm{d}x = \int_0^1 |F^{-1}(u)|\,\mathrm{d}u = \mathbb{E}[|Z|] < +\infty,$$

since Z has finite second moment. Moreover, the Fourier transform of g_{e} is $\mathrm{Re}\{G(\omega)\}$. So, it suffices to reconstruct G, or just $\mathrm{Im}\{G\}$, from $\mathrm{Re}\{G\}$. It is well-known that $\mathrm{Im}\{G\}$ and $\mathrm{Re}\{G\}$ are related by a Hilbert transform under the Titchmarsh analytic conditions [48, Theorem 95], which hold when at least half of Z's distribution is concentrated at one extreme. We also note that the question of determining $\mathrm{Im}\{G\}$ from $\mathrm{Re}\{G\}$ using Hilbert transforms mirrors the classical problem of phase retrieval using minimum phase assumptions (see, e.g., [49]).

C. Antithetic Variates

As a consequence of Theorem 2(b), when a function g is monotonic, $g \circ a$ is monotonic in the opposite direction. This implies that the variance of the antithetic estimator (2) cannot be greater than that of the canonical Monte Carlo estimator with $2n$ samples, $\frac{1}{2n}\sum_{i=1}^{2n} g(Z_i)$. However, this only shows a non-increase in variance, but not any guarantees on the magnitude of the reduction. To establish the latter, we impose additional c-anti-Lipschitz conditions on the pertinent functions to get the following result.

Theorem 4 (Strong antithetic variates)**.** *For any random variable $Z \in \mathbb{R}$ with CDF $F : \mathbb{R} \to [0,1]$, and any monotone increasing c-anti-Lipschitz functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ with constant $c > 0$ such that $\mathbb{E}[f(Z)^2], \mathbb{E}[g(Z)^2] < +\infty$, we have $\mathrm{cov}(f(Z), g(a(Z))) \le -c^2\tau(Z) \le 0$.*

Theorem 4 is proved in Section IV-A. Note that although Theorem 4 does not require Assumption 1, it is required for strong variance reduction results such as Theorem 6. We

also present a classical proof for vanilla antithetic variates in Appendix B.

We next generalize Theorem 4 to present a strengthened version of the antithetic variates method for multivariate strongly isotonic functions.

Theorem 5 (Antithetic variates for strongly isotonic functions)**.** *Consider a random vector $\boldsymbol{X} = (X_1, \dots, X_d) \in \mathbb{R}^d$ where $X_1, \dots, X_d$ are i.i.d. random variables with continuous CDF $F : \mathbb{R} \to [0, 1]$ and antithetic index $\tau > 0$. Then, for any c-strongly isotonic function $g : \mathbb{R}^d \to \mathbb{R}$ with constant $c > 0$ such that $\mathbb{E}[g(\boldsymbol{X})^2] < +\infty$, we have $\operatorname{cov}(g(\boldsymbol{X}), g(\boldsymbol{a}(\boldsymbol{X}))) \le -c^2\tau d$, where $\boldsymbol{a}(\boldsymbol{X}) \triangleq (a(X_1), \dots, a(X_d))$.*

Theorem 5 is proved in Section IV-B. Finally, we further extend Theorem 5 to (output) coordinate-wise isotonic functions and express it as a strong variance reduction inequality.

Theorem 6 (Strong antithetic variance reduction inequality)**.** *Consider a random vector $\boldsymbol{X} = (X_1, \dots, X_d) \in \mathbb{R}^d$ where $X_1, \dots, X_d$ are i.i.d. random variables with continuous CDF $F : \mathbb{R} \to [0, 1]$ and antithetic index $\tau > 0$. Then, for any coordinate-wise c-strongly isotonic function $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^p$ with constant $c > 0$ such that $\mathbb{E}\big[\|\boldsymbol{g}(\boldsymbol{X})\|_2^2\big] < +\infty$, we have*

$$\mathbb{E}\left[\frac{\boldsymbol{g}(\boldsymbol{X}) + \boldsymbol{g}(\boldsymbol{a}(\boldsymbol{X}))}{2}\right] = \mathbb{E}[\boldsymbol{g}(\boldsymbol{X})]$$

and

$$\operatorname{var}\left(\frac{\boldsymbol{g}(\boldsymbol{X}) + \boldsymbol{g}(\boldsymbol{a}(\boldsymbol{X}))}{2}\right) \le \frac{\operatorname{var}(\boldsymbol{g}(\boldsymbol{X})) - c^2\tau p d}{2},$$

where $\boldsymbol{a}(\boldsymbol{X}) = (a(X_1), \dots, a(X_d))$ and the variance of a random vector $\boldsymbol{X}$ is defined as the sum of the variances of each coordinate, i.e., $\operatorname{var}(\boldsymbol{X}) \triangleq \mathbb{E}\big[\|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]\|_2^2\big]$.

Theorem 6 is proved in Section IV-C. It implies that antithetic pairs can be used as "drop-in replacements" for i.i.d. variable pairs to enjoy a reduction in variance. We note that, intuitively, it indeed holds that $\operatorname{var}(g(X)) \ge c^2\tau p d$ in Theorem 6. To see this, suppose we strengthen the coordinate-wise c-strongly isotonic condition on g to the following: $|g_i(x) - g_i(y)| \ge c\,\|x - y\|_2$ for all $i \in \{1, \dots, p\}$ and all $x, y \in \mathbb{R}^d$. Then, for an independent copy X' of X, we have

$$\begin{aligned}
&\mathbb{E}\Big[\|g(X) - \mathbb{E}[g(X)]\|_2^2\Big] = \frac{1}{2}\mathbb{E}\Big[\|g(X) - g(X')\|_2^2\Big] \\
&= \frac{1}{2}\sum_{i=1}^{p}\mathbb{E}\big[(g_i(X) - g_i(X'))^2\big] \ge \frac{c^2 p}{2}\mathbb{E}\Big[\|X - X'\|_2^2\Big] \\
&= c^2 p\,\mathbb{E}\Big[\|X - \mathbb{E}[X]\|_2^2\Big] = c^2 p d \operatorname{var}(X_1) \ge c^2 p d\,\tau(X_1),
\end{aligned}$$

where we use the fact that $X_1, \dots, X_d$ are i.i.d. This shows that the gain in adopting antithetic sampling is at least $c^2 p d(\operatorname{var}(X_1) - \tau(X_1))$.

In the following sections, we explore how our results can improve theoretical guarantees for approximation using Monte Carlo methods in various settings. Note that every random variable considered in the upcoming sections is assumed to be non-atomic with finite second moment.

D. Application to Monte Carlo Integration and Function Approximation

One use of antithetic indices is to compute a lower bound on how much antithetic variates can reduce the variance of a certain Monte Carlo estimation problem without empirical evaluation.

Proposition 7 (Antithetic Monte Carlo integration)**.** *Given a monotone increasing c-anti-Lipschitz function $g : \mathbb{R} \to \mathbb{R}$, the variance of the antithetic Monte Carlo estimator of $\int_0^1 g(t)\,\mathrm{d}t$ satisfies*

$$\operatorname{var}\left(\frac{1}{2n}\sum_{i=1}^{n}(g(U_i) + g(1 - U_i))\right) \le \frac{\operatorname{var}(g(U_1))}{2n} - \frac{c^2}{24n},$$

where $U_1, \dots, U_n \sim \mathsf{Unif}(0, 1)$ are i.i.d. random variables.

Proposition 7 is proved in Section V-A. Next, we apply our results to approximating a function given a stochastic oracle, i.e., estimating $g : \mathbb{R}^p \to \mathbb{R}$, $g(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{Z}}[\phi(\boldsymbol{x}, \boldsymbol{Z})]$ given access to $\phi : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$. Such stochastic functions are commonly encountered in statistics and machine learning contexts. For example, stochastic approximation methods such as the Robbins-Monro algorithm [50] aim to estimate properties of g (e.g., its roots, extrema, etc.) using such a stochastic oracle, while in machine learning contexts, similar stochastic functions are used as random bases to form, e.g., neural network functions, that fit given data [39]. The classic Monte Carlo approach to approximating g at a given point $\boldsymbol{x}$ would be to sample $2n$ values $\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_{2n}$ and evaluate $\frac{1}{2n}\sum_{i=1}^{2n}\phi(\boldsymbol{x}, \boldsymbol{Z}_i)$, which is an unbiased estimator with variance $\operatorname{var}_{\boldsymbol{Z}}(\phi(\boldsymbol{x}, \boldsymbol{Z}))/(2n)$. In the following proposition, we show that the use of antithetic variates can improve the variance of our estimator.

Proposition 8 (Function approximation)**.** *Suppose $g(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{Z}}[\phi(\boldsymbol{x}, \boldsymbol{Z})]$, where $\boldsymbol{Z}$ is a random d-dimensional vector with i.i.d. coordinates that each have continuous CDF F and antithetic index $\tau > 0$, and $\phi : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$ is a function with c-strongly isotonic $\phi(\boldsymbol{x}, \cdot)$ for all $\boldsymbol{x}$. Then, given a point $\boldsymbol{x}$, the antithetic estimator*

$$\hat{g}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\phi(\boldsymbol{x}, \boldsymbol{Z}_i) + \phi(\boldsymbol{x}, \boldsymbol{a}(\boldsymbol{Z}_i))}{2}$$

is an unbiased estimator of $g(\boldsymbol{x})$ with a variance of

$$\mathbb{E}\left[(\hat{g}(\boldsymbol{x}) - g(\boldsymbol{x}))^2\right] \le \frac{\operatorname{var}_{\boldsymbol{Z}}(\phi(\boldsymbol{x}, \boldsymbol{Z})) - c^2\tau d}{2n}.$$

Moreover, if ϕ is L-Lipschitz under ℓ^∞-norm with respect to $\boldsymbol{Z}$ for all $\boldsymbol{x}$ (with $L > c$), we also have

$$\mathbb{E}\left[(\hat{g}(\boldsymbol{x}) - g(\boldsymbol{x}))^2\right] \le \frac{d}{2n}(L^2\operatorname{var}(Z) - c^2\tau),$$

where $\operatorname{var}(Z)$ is the variance of a single component of $\boldsymbol{Z}$. If we further assume that the domain of $\boldsymbol{x}$ is restricted to a compact set $\mathcal{X} \subset \mathbb{R}^p$ with volume $\operatorname{vol}(\mathcal{X})$ and that F is symmetric,

$$\mathbb{E}\left[\|\hat{g} - g\|_2\right] \le \sqrt{\frac{d}{2n}\operatorname{var}(Z)\operatorname{vol}(\mathcal{X})(L^2 - c^2)},$$

where $\|\cdot\|_2$ denotes the $\mathcal{L}^2$-norm when applied to functions.

Proposition 8 is proved in Section V-B. We briefly mention an interesting special case of the problem of estimating a stochastic function g, which can be recast in integral transform (or "basis expansion") form $g(\boldsymbol{x}) = \int_{\mathbb{R}^d} \phi(\boldsymbol{x}, \boldsymbol{z}) p(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{z}$, where $p(\boldsymbol{z})$ is the PDF of $\boldsymbol{Z}$. Consider the case where $\phi : [0, a] \times I \to \mathbb{R}$, $\phi(x, z) = e^{-xz}$ (for some $a > 0$ and compact interval I) so that $g(x)$ is the Laplace transform of $p(z)$. Under mild conditions, the Hausdorff-Bernstein-Widder theorem states that g is such a Laplace transform if and only if g is completely monotone (i.e., has alternating signed derivatives) [51], [52]. Since the "Laplace basis" $\phi(x, z)$ is (uniformly) strongly isotonic with respect to z if $x \in [0, a]$ and $z \in I$, one special case of the function class we are approximating is a non-parametric class of completely monotone functions. Such functions are known to characterize transfer functions of externally positive linear systems in signal processing and control applications [53].

E. Application to Concentration of Measure

Our bounds can also be used to tighten many theoretical results in probability that depend on the variance of samples. We present the following improvement to Bernstein's inequality [17] as an example.

Proposition 9 (Bernstein's inequality for antithetic variates)**.** *Let $X_1, \dots, X_n$ be i.i.d. zero-mean real-valued random variables with variance σ^2, continuous CDF F such that $F(-K) = 1 - F(K) = 0$ for some $K > 0$, and antithetic index τ. Let $X_{n+1}, \dots, X_{2n}$ be their antithetic counterparts. Then, for all $t \geq 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^{2n} X_i\right| \geq t\right) \leq 2 \exp\left(-\frac{t^2/2}{2n(\sigma^2 - \tau) + 2Kt/3}\right).$$

Proposition 9 is proved in Section VI-A. Note that the bound is asymptotically tighter than the usual Bernstein's inequality for $2n$ i.i.d. variables,

$$\mathbb{P}\left(\left|\sum_{i=1}^{2n} X_i\right| \geq t\right) \leq 2 \exp\left(-\frac{t^2/2}{2n\sigma^2 + Kt/3}\right),$$

when t/n is dominated by σ^2/K as is the case in the sub-Gaussian regime [17]. Although this is an improvement in an asymptotic sense, the $2Kt/3$ term prevents it from being unconditionally superior. This is due to a difficulty in bounding the sum $X + a(X)$. Without any additional assumptions, it is difficult to improve on the trivial bound: $|X| \leq K$ a.s. $\Rightarrow$ $|X + a(X)| \leq 2K$ a.s. This is especially troublesome for inequalities such as Bennett's inequality which have a greater dependence on K. Fortunately, it is possible to improve this bound with an additional (mild) assumption on the distribution of X.

Proposition 10 (Condition for tighter bound on antithetic sum)**.** *Let X be a random variable with CDF F such that $|X| \leq K$ a.s. for some constant K. Suppose there exist some random variables Y and Z with symmetric distributions such that $\mathbb{E}[Y] = -K/2$, $\mathbb{E}[Z] = K/2$, $-K \leq Y \leq 0$ a.s., $0 \leq Z \leq K$ a.s., and $Y \preceq X \preceq Z$, where $A \preceq B$ means A has (first-order) stochastic dominance over B, i.e., $\mathbb{P}(A \geq x) \geq \mathbb{P}(B \geq x)$ for all $x \in \mathbb{R}$ [54]. Then, $|X + a(X)| \leq K$ a.s.*

Furthermore, if such Y and Z exist, then $F(K/2) \geq 1/2$ and $F(-K/2) \leq 1/2$. Conversely, if $F(0) = 1/2$ (i.e., X has median 0), or F is convex in $[-K, 0]$ and concave in $[0, K]$, or F is $1/K$-Lipschitz, then such Y and Z exist.

Proposition 10 is proved in Section VI-B. As the various sufficient conditions imply, the condition described in Proposition 10 is very mild. The stochastic dominance condition $Y \preceq X \preceq Z$ holds for many common distributions such as the beta distribution and triangular distribution (when shifted to have zero mean). Under Proposition 10's assumptions, concentration inequalities such as Bernstein's inequality and Bennett's inequality can be improved using antithetic variates without any tradeoff.

Proposition 11 (Improved Bernstein's inequality for antithetic variates)**.** *In addition to the assumptions of Proposition 9, further assume that Proposition 10 holds. Then, for all $t \geq 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^{2n} X_i\right| \geq t\right) \leq 2 \exp\left(-\frac{t^2/2}{2n(\sigma^2 - \tau) + Kt/3}\right).$$

Furthermore, the improved bound on $X + a(X)$ allows for a direct improvement to Bennett's inequality as well.

Proposition 12 (Bennett's inequality for antithetic variates)**.** *Let $X_1, \dots, X_n$ be i.i.d. zero-mean random variables with variance σ^2, continuous CDF F such that $F(-K) = 1 - F(K) = 0$ for some $K > 0$, and antithetic index τ. Let $X_{n+1}, \dots, X_{2n}$ be their antithetic counterparts. Suppose Proposition 10 holds. Then, for all $t \geq 0$,*

$$\mathbb{P}\left(\sum_{i=1}^{2n} X_i \geq t\right) \leq \exp\left(-\frac{2n(\sigma^2 - \tau)}{K^2} h\left(\frac{tK}{2n(\sigma^2 - \tau)}\right)\right),$$

where $h(x) = (1 + x)\log(1 + x) - x$.

Propositions 11 and 12 are immediate consequences of Proposition 10 and the standard inequalities in [17]. Next, we shift our attention to concentration inequalities regarding Gaussian variables. The symmetry of Gaussian distributions can be exploited to derive an improvement to a classical concentration inequality regarding Lipschitz functions.

Proposition 13 (Antithetic Lipschitz concentration)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-Lipschitz (under ℓ^2-norm) and c-strongly isotonic function. Let $\boldsymbol{X} = (X_1, \dots, X_n)$ be i.i.d. $\mathcal{N}(0, 1)$ with antithetic map $\boldsymbol{a}(\boldsymbol{x}) = -\boldsymbol{x}$. Let $g(\boldsymbol{x}) = \frac{f(\boldsymbol{x}) + f(\boldsymbol{a}(\boldsymbol{x}))}{\sqrt{2}} = \frac{f(\boldsymbol{x}) + f(-\boldsymbol{x})}{\sqrt{2}}$. Then, for all $t \geq 0$,*

$$\mathbb{P}\left(|g(\boldsymbol{X}) - \mathbb{E}[g(\boldsymbol{X})]| \geq t\right) \leq 2 \exp\left(-\frac{t^2}{2(L - c)^2}\right).$$

Proposition 13 is proved in Section VI-C. It improves on the standard Gaussian concentration inequality for Lipschitz functions [55]: $\mathbb{P}(|f(\boldsymbol{X}) - \mathbb{E}[f(\boldsymbol{X})]| \geq t) \leq 2\exp(-t^2/(2L^2))$. The normalizing constant $\sqrt{2}$ might seem unusual, but it naturally arises from the ℓ^2-norm Lipschitzness assumption. Suppose that we define another version of g, $g' : \mathbb{R}^{2n} \to \mathbb{R}$, that uses $2n$ i.i.d. variables instead of using antithetic variates:

$g'(\boldsymbol{x}) = f((x_1, \ldots, x_n)) + f((x_{n+1}, \ldots, x_{2n}))$. Then, g' will be $\sqrt{2}L$-Lipschitz, so g' must be scaled by $1/\sqrt{2}$ to maintain the same Lipschitz constant.

F. Application to Stochastic Optimization

Next, we proceed to an application of our antithetic variance inequalities in the domain of first-order optimization theory. In machine learning problems where the objective function can be expressed as the average of some loss function evaluated at each data point (e.g., empirical risk minimization), stochastic gradient descent (SGD) is often used instead of plain gradient descent (GD) to reduce the cost of evaluating the gradient when the dataset is large. Although the stochastic gradient is an unbiased estimate of the full gradient, its high variance could sometimes make SGD converge more slowly than GD.

Thus, many techniques have been developed to reduce the variance of the stochastic gradient while keeping it unbiased [18], [29]–[33], [35]–[37], [56]. The use of antithetic variates for this purpose has been studied by [19], [28], but theoretical results that indicate an improvement from vanilla SGD do not exist. Here, we outline an alternative method of utilizing antithetic variates and show it to have better worst-case convergence bounds than SGD.

Given an L-smooth (i.e., continuously differentiable with L-Lipschitz gradient) but possibly non-convex function $f : \mathbb{R}^p \to \mathbb{R}$, suppose $\boldsymbol{g} : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}^p$ is an unbiased stochastic first-order oracle (SFO) of f, i.e., $\mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{Z})] = \nabla f(\boldsymbol{x})$ and $\operatorname{var}_{\boldsymbol{Z}}(\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{Z})) \le \sigma^2$ for all $\boldsymbol{x} \in \mathbb{R}^p$, where $\boldsymbol{Z} = (Z_1, \ldots, Z_d)$ and $Z_1, \ldots, Z_d$ are i.i.d. random variables with known distribution and antithetic index $\tau > 0$. Even though we assume that a global optimum $\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x} \in \mathbb{R}^p} f(\boldsymbol{x})$ exists, finding it for non-convex f may be NP-hard (see [57]). Instead, we aim to find an ε-stationary point, i.e., a random vector $\hat{\boldsymbol{x}} \in \mathbb{R}^p$ such that $\mathbb{E}\left[\|\nabla \boldsymbol{f}(\hat{\boldsymbol{x}})\|_2^2\right] \le \varepsilon$ for some desired accuracy $\varepsilon > 0$. We note that most of the above, e.g., L-smoothness, SFO, etc., are standard assumptions in the optimization literature.

Starting from some $\boldsymbol{x}_1 \in \mathbb{R}^p$, the *antithetic SGD* algorithm repeats the following iteration:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \frac{\boldsymbol{g}(\boldsymbol{x}_t, \boldsymbol{Z}_t) + \boldsymbol{g}(\boldsymbol{x}_t, \boldsymbol{a}(\boldsymbol{Z}_t))}{2}, \qquad \boldsymbol{Z}_t \sim P_Z^d,$$

where $\eta > 0$ is the step-size and $\boldsymbol{a}$ is the multidimensional antithetic map as defined in Theorem 6.

Theorem 14 (Antithetic SGD in the non-convex setting)**.** *Suppose $\boldsymbol{g}$ is c-strongly isotonic with respect to $\boldsymbol{Z}$ for all $\boldsymbol{x}$. After T iterations, let $\hat{\boldsymbol{x}}$ be a random vector such that $\mathbb{P}(\hat{\boldsymbol{x}} = \boldsymbol{x}_t) = 1/T$ for $t \in \{1, \ldots, T\}$. Then, the antithetic SGD method with $\eta \le 1/L$ satisfies*

$$\mathbb{E}_{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_T, \hat{\boldsymbol{x}}}\left[\|\nabla f(\hat{\boldsymbol{x}})\|_2^2\right] \le \frac{2\Delta_1}{T\eta} + \frac{L\eta(\sigma^2 - c^2\tau p d)}{2},$$

where $\Delta_1 = f(\boldsymbol{x}_1) - f(\boldsymbol{x}^*)$.

Theorem 14 is proved in Section VII-A. Note that an equivalent (mini-batch) SGD implementation that takes two i.i.d. samples per iteration has the bound

$$\mathbb{E}_{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_T, \hat{\boldsymbol{x}}}\left[\|\nabla f(\hat{\boldsymbol{x}})\|_2^2\right] \le \frac{2\Delta_1}{T\eta} + \frac{L\eta\sigma^2}{2},$$

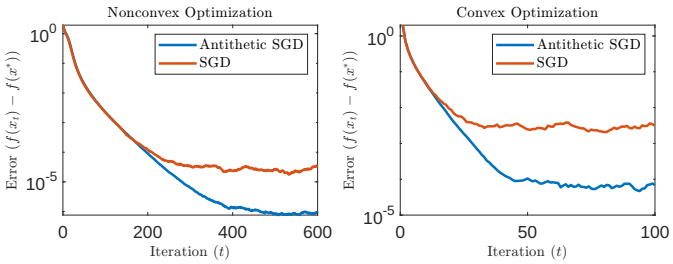

Fig. 1. A comparison between the performance of SGD and antithetic SGD when used to optimize nonconvex (left) and strongly convex (right) functions.

which lacks the $c^2\tau p d$ term. Similarly, we can derive improved guarantees in the strongly convex setting, where f is μ-strongly convex, i.e., $f(\boldsymbol{y}) - f(\boldsymbol{x}) \ge \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$.

Theorem 15 (Antithetic SGD in the strongly convex setting)**.** *In addition to the assumptions of Theorem 14, further assume that f is μ-strongly convex. If the step-size satisfies $\eta \le \mu/L^2$,*

$$\mathbb{E}_{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_T}\left[\|\boldsymbol{x}_{T+1} - \boldsymbol{x}^*\|_2^2\right] \le (1 - \eta\mu)^T \delta_1 + \eta \frac{\sigma^2 - c^2\tau p d}{2\mu}$$

where $\delta_1 = \|\boldsymbol{x}_1 - \boldsymbol{x}^*\|_2^2$.

Theorem 15 is proved in Section VII-B. Figure 1 shows examples of two experiments that empirically demonstrate the improved performance of antithetic SGD as implied in Theorems 14 and 15. The experiments compare the performance of constant step-size SGD (with batches of two samples per iteration so that its rate of oracle calls is equivalent to that of antithetic SGD) and antithetic SGD when used to find the minima of $f_1(x) = 1 - \frac{x}{1+x^2} + \frac{\log(1+x)}{2x}$ (a non-convex function) and $f_2(x) = x^4 + x^2 + 1$ (a strongly convex function) using the first-order oracles $g_1(x, Z) = \frac{x^2-1}{(1+x^2)^2} - \frac{Z^3}{(1+xZ^2)^2}$ and $g_2(x, Z) = z(z + 4/3)(x^2 + 1) - (x - 1)^2 + 4x^3$, respectively, where $Z \sim \mathsf{Unif}(0, 1)$. As predicted by Theorems 14 and 15, antithetic SGD achieves a lower noise floor than SGD.

III. Proofs of Properties of Antithetic Maps and Indices

A. Proof of Theorem 2

Proof.

Part (a): From the definition of the antithetic map, we have

$$\begin{aligned}
&\mathbb{P}(a_Z(Z) \le x) = \mathbb{P}(F^{-1}(1 - F(Z)) \le x) \\
&\overset{(a)}{=} \mathbb{P}(1 - F(Z) \le F(x)) \overset{(b)}{=} \mathbb{P}(Z \ge F^{-1}(1 - F(x))) \\
&= 1 - F(F^{-1}(1 - F(x))) \overset{(c)}{=} F(x),
\end{aligned}$$

where (a) and (b) hold due to $F(x) \ge y \Leftrightarrow x \ge F^{-1}(y)$ [42] and (c) follows from $F(F^{-1}(y)) = y$ when Z is non-atomic (see note below Definition 1).

Part (b): Since F^{-1} is non-decreasing [42] and $1 - F$ is non-increasing, the composition $F^{-1}(1 - F(\cdot))$ is non-increasing. Furthermore, if F is continuous and strictly increasing, F^{-1} is continuous and strictly increasing in the support of Z [42], thus a_Z is also strictly decreasing and continuous.

Part (c): Consider any strictly decreasing and continuous $f:\mathbb{R}\to\mathbb{R}$ such that Z and $f(Z)$ are identically distributed. Then, for any $t\in\mathbb{R}$, we have

$$F(t)=\mathbb{P}(f(Z)\le t)=\mathbb{P}(Z\ge f^{-1}(t))\overset{(a)}{=}1-F(f^{-1}(t)),$$

where (a) uses the continuity of F. This implies that $t=F^{-1}(1-F(f^{-1}(t)))$ for all $t\in\mathbb{R}$. Hence, the inverse of f^{-1} is the strictly decreasing function $f(t)=F^{-1}(1-F(t))$. Therefore, $f(t)=F^{-1}(1-F(t))$ is the unique strictly decreasing and continuous function such that Z and $f(Z)$ are identically distributed.

Part (d): $Y=\alpha Z+\beta$ has CDF $F_Y(y)=F((y-\beta)/\alpha)$ and inverse CDF $F_Y^{-1}(t)=\alpha F^{-1}(t)+\beta$. Thus,

$$\begin{aligned}a_Y(y)&=F_Y^{-1}(1-F_Y(y))\\&=\alpha F^{-1}\left(1-F\left(\frac{y-\beta}{\alpha}\right)\right)+\beta=\alpha a_Z\left(\frac{y-\beta}{\alpha}\right)+\beta.\end{aligned}$$

In addition,

$$\begin{aligned}\tau(Y)&=-\mathrm{cov}(Y,a_Y(Y))=-\mathrm{cov}(\alpha Z+\beta,\alpha a_Z(Z)+\beta)\\&=-\alpha^2\mathrm{cov}(Z,a_Z(Z))=\alpha^2\tau(Z).\end{aligned}$$

Part (e): $\tau(Z)=-\mathrm{cov}(Z,a_Z(Z))\ge 0$ follows from Chebyshev's association inequality [17] due to a_Z being non-increasing from Theorem 2(b). It can be shown that Z_t has variance $\mathrm{var}(Z_t)=\sigma^2$ and antithetic index

$$\tau(Z_t)=\frac{(1-t)(10t^2-6t+1)}{2t^3}\sigma^2,$$

thus $\lim_{t\to 1}\tau(Z_t)=0$.

Part (f): Per the Cauchy-Schwarz inequality,

$$|\mathrm{cov}(Z,a_Z(Z))|^2\le \mathrm{var}(Z)\mathrm{var}(a_Z(Z))=\mathrm{var}^2(Z),$$

with equality if and only if $a_Z(Z)=cZ+d$ a.s. for some constants c and d. Since a_Z is non-increasing and $a_Z(Z)$ and Z are identically distributed due to Theorem 2(a) and Theorem 2(b), it can be inferred that $c=-1$ and Z must be symmetric in order for the equality to hold.

Part (g): The upper bound follows immediately from Popoviciu's inequality [58], which states that $\mathrm{var}(Z)\le L^2/4$, and Theorem 2(f). Since $\mathrm{var}(B_k)=k/(8+4k)$, using Theorem 2(d) and Theorem 2(f) we have that $\tau(Z_k)=L^2k/(8+4k)$. Thus, $\lim_{k\to\infty}\tau(Z_k)=L^2/4$, achieving the supremum. □

B. Proof of Theorem 3

Proof.

Part (a): We begin by noting that the ℓ^2-optimal transport map from P to Q on the real line is known to be the monotone rearrangement (see, e.g., [47, Corollary 4.6]), i.e., $F_Q^{-1}\circ F_P$ where F_P and F_Q are the CDFs of P and Q, respectively.

Next, set P and Q such that $Z\sim P$ and $-Z\sim Q$. Then, $F_Q(x)=1-F_P(-x)$, thus $F_Q^{-1}(t)=-F_P^{-1}(1-t)$. Therefore, the monotone rearrangement from P to Q is $F_Q^{-1}(F_P(\cdot))=-F_P^{-1}(1-F_P(\cdot))=-a_Z$. It is well-known and easy to prove that the maximum correlation problem is equivalent to the optimal transport problem with ℓ^2-cost (see, e.g., [59]), completing the proof.

Part (b): In one dimension, the 2-Wasserstein distance is known to be $W_2(P,Q)=\sqrt{\int_0^1(F_P^{-1}(t)-F_Q^{-1}(t))^2\,\mathrm{d}t}$ (see, e.g., [59]). If $Z\sim P$ and $2\mathbb{E}[Z]-Z\sim Q$, we have

$$\begin{aligned}F_Q(x)&=\mathbb{P}(2\mathbb{E}[X]-X\le x)=\mathbb{P}(X\ge 2\mathbb{E}[X]-x)\\&=1-F_P(2\mathbb{E}[X]-x)\end{aligned}$$

and $F_Q^{-1}(t)=2\mathbb{E}[X]-F_P^{-1}(1-t)$. Thus,

$$\begin{aligned}W_2(P,Q)&=\sqrt{\int_0^1(F_P^{-1}(t)-F_Q^{-1}(t))^2\,\mathrm{d}t}\\&=\sqrt{\int_0^1(F_P^{-1}(t)-2\mathbb{E}[Z]+F_P^{-1}(1-t))^2\,\mathrm{d}t}\\&=\sqrt{\int_{\mathbb{R}}(z-2\mathbb{E}[Z]+F_P^{-1}(1-F_P(z)))^2\,\mathrm{d}P(z)}\\&=\sqrt{\int_{\mathbb{R}}(z+a_Z(z)-2\mathbb{E}[Z])^2\,\mathrm{d}P(z)}\\&\overset{(a)}{=}\sqrt{\mathrm{var}(Z+a_Z(Z))}=\sqrt{2\mathrm{var}(Z)-2\tau(Z)},\end{aligned}$$

where (a) follows from $\mathbb{E}[Z+a_Z(Z)]=2\mathbb{E}[Z]$. We also note that the notation $\int\cdot\,\mathrm{d}P(z)$ refers to a Lebesgue integral with respect to the measure defined by P in the sequel.

Part (c): Consider the square of W_2 distance which we will upper bound by using the identity coupling:

$$\begin{aligned}W_2(P,R)^2&=\inf_{\pi\in\Pi(P,R)}\int_{(x,y)}(x-y)^2\pi(x,y)\,\mathrm{d}x\,\mathrm{d}y\\&\le\int_{\mathbb{R}}\left(z-\frac{z+a_Z(z)}{2}\right)^2\mathrm{d}P(z)\\&=\mathbb{E}_{Z\sim P}\left[\left(\frac{Z-a_Z(Z)}{2}\right)^2\right]\\&=\frac{1}{4}\mathbb{E}[Z^2]+\frac{1}{4}\mathbb{E}[a_Z(Z)^2]-\frac{1}{2}\mathbb{E}[Za_Z(Z)]\\&=\frac{\mathbb{E}[Z^2]-\mathbb{E}[Za_Z(Z)]}{2}=\frac{\mathrm{var}(Z)+\tau(Z)}{2}.\end{aligned}$$

Part (d): Similarly to (c),

$$\begin{aligned}W_2(P,S)^2&=\inf_{\pi\in\Pi(P,S)}\int_{(x,y)}(x-y)^2\pi(x,y)\,\mathrm{d}x\,\mathrm{d}y\\&\le\mathbb{E}_{Z\sim P}\left[\left(\frac{Z+a_Z(Z)}{2}-\mathbb{E}[Z]\right)^2\right]\\&=\mathrm{var}\left(\frac{Z+a_Z(Z)}{2}\right)=\frac{\mathrm{var}(Z)-\tau(Z)}{2}.\end{aligned}$$

This completes the proof. □

IV. Proofs of Antithetic Variates Results

A. Proof of Theorem 4

Proof. We begin by letting $\tilde{Z}$ denote an independent copy of Z. Then, we have

$$\begin{aligned}&\mathbb{E}\Big[\Big(f(Z)-f(\tilde{Z})\Big)\Big(g(a(\tilde{Z}))-g(a(Z))\Big)\Big]\\&\overset{(a)}{\ge}c^2\mathbb{E}\Big[(Z-\tilde{Z})\Big(a(\tilde{Z})-a(Z)\Big)\Big]\\&\overset{(b)}{=}2c^2(\mathbb{E}[Z]\mathbb{E}[a(Z)]-\mathbb{E}[Za(Z)])=-2c^2\mathrm{cov}(Z,a(Z)),\end{aligned}$$

where (a) follows from the monotonicity and anti-Lipschitzness of f and g and the fact that a is monotone non-increasing, and (b) uses the fact that $\tilde{Z}$ is an independent copy of Z. Furthermore, note that

$$\begin{aligned}&\mathbb{E}\Big[\Big(f(Z)-f(\tilde{Z})\Big)\Big(g(a(\tilde{Z}))-g(a(Z))\Big)\Big]\\&=2\left(\mathbb{E}[f(Z)]\,\mathbb{E}\Big[g(a(\tilde{Z}))\Big]-\mathbb{E}[f(Z)g(a(Z))]\right)\\&=-2\operatorname{cov}(f(Z),g(a(Z)))\,,\end{aligned}$$

where we again use the fact that $\tilde{Z}$ is an independent copy of Z. Together, these inequalities yield the desired bound $\operatorname{cov}(f(Z),g(a(Z))) \le c^2\operatorname{cov}(Z,a(Z)) = -c^2\tau(Z) \le 0$, where the nonpositivity follows from Theorem 2(e). □

B. Proof of Theorem 5

Proof. We establish this result by induction on the dimension d. When $d=1$, the result holds due to Theorem 4. Suppose the result holds for any arbitrary fixed d. We will use this inductive hypothesis to show that the result holds for $d+1$.

To this end, consider $d+1$ i.i.d. random variables $X_1,\dots,X_{d+1}$ with continuous CDF $F:\mathbb{R}\to[0,1]$ and any c-strongly isotonic function $g:\mathbb{R}^{d+1}\to\mathbb{R}$. Let $\boldsymbol{W}=(X_2,\dots,X_{d+1})$ and $\boldsymbol{X}=(X_1,\boldsymbol{W})=(X_1,\dots,X_{d+1})$. Then,

$$\begin{aligned}&\mathbb{E}[g(\boldsymbol{X})g(\boldsymbol{a}(\boldsymbol{X}))]\overset{\text{(a)}}{=}\mathbb{E}_{\boldsymbol{W}}[\mathbb{E}_{X_1}[g(X_1,\boldsymbol{W})g(a(X_1),\boldsymbol{a}(\boldsymbol{W}))|\boldsymbol{W}]]\\&\overset{\text{(b)}}{=}\int_{\mathbb{R}^d}\mathbb{E}_{X_1}[g(X_1,\boldsymbol{w})g(a(X_1),\boldsymbol{a}(\boldsymbol{w}))]\,\mathrm{d}P(\boldsymbol{w})\\&\overset{\text{(c)}}{\le}\int_{\mathbb{R}^d}\mathbb{E}_{X_1}[g(X_1,\boldsymbol{w})]\,\mathbb{E}_{X_1}[g(a(X_1),\boldsymbol{a}(\boldsymbol{w}))]\,\mathrm{d}P(\boldsymbol{w})-c^2\tau\\&=\mathbb{E}_{\boldsymbol{W}}[\mathbb{E}_{X_1}[g(X_1,\boldsymbol{W})|\boldsymbol{W}]\,\mathbb{E}_{X_1}[g(a(X_1),\boldsymbol{a}(\boldsymbol{W}))|\boldsymbol{W}]]-c^2\tau\\&\overset{\text{(d)}}{\le}\mathbb{E}_{\boldsymbol{W}}[\mathbb{E}_{X_1}[g(X_1,\boldsymbol{W})|\boldsymbol{W}]]\,\mathbb{E}_{\boldsymbol{W}}[\mathbb{E}_{X_1}[g(a(X_1),\boldsymbol{a}(\boldsymbol{W}))|\boldsymbol{W}]]\\&\qquad-c^2\tau(d+1)\\&\overset{\text{(e)}}{=}\mathbb{E}[g(\boldsymbol{X})]\,\mathbb{E}[g(\boldsymbol{a}(\boldsymbol{X}))]-c^2\tau(d+1),\end{aligned}$$

where $\boldsymbol{W}\sim P$, (a) and (e) follow from the tower property, (b) uses the independence of X_1 and $\boldsymbol{W}$, (c) follows from Theorem 4 because $h(t)=g(t,\boldsymbol{w})$ and $h'(t)=g(t,\boldsymbol{a}(\boldsymbol{w}))$ are monotone non-decreasing and c-anti-Lipschitz, and (d) follows from the inductive hypothesis and the independence of X_1 and $\boldsymbol{W}$ because $h(\boldsymbol{w})=\mathbb{E}_{X_1}[g(X_1,\boldsymbol{w})]$ is c-strongly isotonic and $\mathbb{E}_{X_1}[g(a(X_1),\boldsymbol{a}(\boldsymbol{W}))|\boldsymbol{W}=\boldsymbol{w}]=h(\boldsymbol{a}(\boldsymbol{w}))$ due to Theorem 2(a). Therefore, we have established by induction that $\operatorname{cov}(g(\boldsymbol{X}),g(\boldsymbol{a}(\boldsymbol{X})))\le -c^2\tau d$, completing the proof. □

C. Proof of Theorem 6

Proof. The first result follows immediately from Theorem 2(a). Next, observe that

$$\begin{aligned}&\operatorname{var}\left(\frac{\boldsymbol{g}(\boldsymbol{X})+\boldsymbol{g}(\boldsymbol{a}(\boldsymbol{X}))}{2}\right)\\&=\mathbb{E}\left[\left\|\frac{\boldsymbol{g}(\boldsymbol{X})+\boldsymbol{g}(\boldsymbol{a}(\boldsymbol{X}))}{2}-\mathbb{E}\left[\frac{\boldsymbol{g}(\boldsymbol{X})+\boldsymbol{g}(\boldsymbol{a}(\boldsymbol{X}))}{2}\right]\right\|_2^2\right]\\&=\frac{1}{4}\mathbb{E}\Big[\|\boldsymbol{g}(\boldsymbol{X})-\mathbb{E}[\boldsymbol{g}(\boldsymbol{X})]\|_2^2\Big]+\frac{1}{4}\mathbb{E}\Big[\|\boldsymbol{g}(\boldsymbol{a}(\boldsymbol{X}))-\mathbb{E}[\boldsymbol{g}(\boldsymbol{a}(\boldsymbol{X}))]\|_2^2\Big]\\&\quad+\frac{1}{2}\mathbb{E}\Big[(\boldsymbol{g}(\boldsymbol{X})-\mathbb{E}[\boldsymbol{g}(\boldsymbol{X})])^{\mathrm{T}}(\boldsymbol{g}(\boldsymbol{a}(\boldsymbol{X}))-\mathbb{E}[\boldsymbol{g}(\boldsymbol{a}(\boldsymbol{X}))])\Big]\\&\overset{\text{(a)}}{=}\frac{\operatorname{var}(\boldsymbol{g}(\boldsymbol{X}))}{2}+\frac{1}{2}\sum_{i=1}^{p}\operatorname{cov}(g_i(\boldsymbol{X}),g_i(\boldsymbol{a}(\boldsymbol{X})))\\&\overset{\text{(b)}}{\le}\frac{\operatorname{var}(\boldsymbol{g}(\boldsymbol{X}))-c^2\tau pd}{2},\end{aligned}$$

where (a) uses Theorem 2(a) and (b) follows from Theorem 5, completing the proof. □

V. Proofs of Monte Carlo Integration and Function Approximation Results

A. Proof of Proposition 7

Proof. Let $h(\boldsymbol{U})=\frac{1}{n}\sum_{i=1}^n g(U_i)$. Since h is c/n-strongly isotonic, Theorem 6 implies

$$\begin{aligned}&\operatorname{var}\left(\frac{1}{2n}\sum_{i=1}^{n}(g(U_i)+g(1-U_i))\right)=\operatorname{var}\left(\frac{h(\boldsymbol{U})+h(\boldsymbol{a}(\boldsymbol{U}))}{2}\right)\\&\le\frac{\operatorname{var}(g(\boldsymbol{U}))-\left(\frac{c}{n}\right)^2\tau n}{2}\overset{\text{(a)}}{=}\frac{\operatorname{var}(g(U_1))}{2n}-\frac{c^2}{24n},\end{aligned}$$

where (a) follows from $\tau(U_i)=1/12$. □

B. Proof of Proposition 8

Proof. The proof follows from adapting arguments in [20], [28]. Using Theorem 5 and the fact that $\phi(\boldsymbol{x},\cdot):\mathbb{R}^d\to\mathbb{R}$ is c-strongly isotonic for all $\boldsymbol{x}$, we have $\operatorname{cov}_{\boldsymbol{Z}}(\phi(\boldsymbol{x},\boldsymbol{Z}),\phi(\boldsymbol{x},\boldsymbol{a}(\boldsymbol{Z})))\le -c^2\tau d$ for all $\boldsymbol{x}$. Thus,

$$\begin{aligned}&\mathbb{E}\left[(\hat{g}(\boldsymbol{x})-g(\boldsymbol{x}))^2\right]\\&=\frac{\operatorname{var}_{\boldsymbol{Z}}(\phi(\boldsymbol{x},\boldsymbol{Z}))+\operatorname{cov}_{\boldsymbol{Z}}(\phi(\boldsymbol{x},\boldsymbol{Z}),\phi(\boldsymbol{x},\boldsymbol{a}(\boldsymbol{Z})))}{2n}\\&\le\frac{\operatorname{var}_{\boldsymbol{Z}}(\phi(\boldsymbol{x},\boldsymbol{Z}))-c^2\tau d}{2n},\end{aligned}$$

where $\hat{g}(\boldsymbol{x})=\frac{1}{n}\sum_{i=1}^n\frac{\phi(\boldsymbol{x},\boldsymbol{Z}_i)+\phi(\boldsymbol{x},\boldsymbol{a}(\boldsymbol{Z}_i))}{2}$ is an unbiased estimator of $g(\boldsymbol{x})$. In addition, if $\phi(\boldsymbol{x},\cdot):\mathbb{R}^d\to\mathbb{R}$ is L-Lipschitz with respect to the ℓ^∞-norm for all $\boldsymbol{x}$,

$$\begin{aligned}&\operatorname{var}(\phi(\boldsymbol{x},\boldsymbol{Z}))=\frac{1}{2}\operatorname{var}(\phi(\boldsymbol{x},\boldsymbol{Z})-\phi(\boldsymbol{x},\boldsymbol{Z}'))\\&=\frac{1}{2}\mathbb{E}[(\phi(\boldsymbol{x},\boldsymbol{Z})-\phi(\boldsymbol{x},\boldsymbol{Z}'))^2]\overset{\text{(a)}}{\le}\frac{L^2}{2}\mathbb{E}[\|\boldsymbol{Z}-\boldsymbol{Z}'\|_\infty^2]\\&\overset{\text{(b)}}{\le}\frac{L^2}{2}\mathbb{E}[\|\boldsymbol{Z}-\boldsymbol{Z}'\|_2^2]=L^2d\operatorname{var}(Z),\end{aligned}$$

where $\boldsymbol{Z}'$ is an independent copy of $\boldsymbol{Z}$, (a) is due to Lipschitz continuity, (b) is due to the monotonicity of ℓ^p-norms, and $\operatorname{var}(Z)$ is the variance of a single component of $\boldsymbol{Z}$. Thus,

$$\mathbb{E}\left[(\hat{g}(\boldsymbol{x})-g(\boldsymbol{x}))^2\right]\le\frac{d}{2n}(L^2\operatorname{var}(Z)-c^2\tau).$$

When the distribution of Z is symmetric, this simplifies to $\mathbb{E}\left[(\hat{g}(\boldsymbol{x})-g(\boldsymbol{x}))^2\right]\le\frac{d}{2n}\operatorname{var}(Z)(L^2-c^2)$. Furthermore, if the domain of $\boldsymbol{x}$ is a compact set $\mathcal{X}$ with volume $\operatorname{vol}(\mathcal{X})>0$,

$$\begin{aligned}&\mathbb{E}\big[\|\hat{g}-g\|_2\big]=\mathbb{E}\left[\sqrt{\int_{\mathcal{X}}(\hat{g}(\boldsymbol{x})-g(\boldsymbol{x}))^2d\boldsymbol{x}}\right]\\&\overset{\text{(a)}}{\le}\sqrt{\mathbb{E}\left[\int_{\mathcal{X}}\hat{g}(\boldsymbol{x})-g(\boldsymbol{x}))^2d\boldsymbol{x}\right]}\overset{\text{(b)}}{=}\sqrt{\int_{\mathcal{X}}\mathbb{E}[(\hat{g}(\boldsymbol{x})-g(\boldsymbol{x}))^2]\,d\boldsymbol{x}}\\&\le\sqrt{\int_{\mathcal{X}}\frac{\operatorname{dvar}(Z)(L^2-c^2)d}{2n}\boldsymbol{x}}=\sqrt{\frac{\operatorname{dvar}(Z)\operatorname{vol}(\mathcal{X})(L^2-c^2)}{2n}},\end{aligned}$$

where (a) is due to Jensen's inequality and (b) follows from Tonelli's theorem. This proves the bound on $\mathbb{E}\big[\|\hat{g}-g\|_2\big]$. □

VI. Proofs of Concentration of Measure Results

A. Proof of Proposition 9

Proof. Recall that for $Y_1,\dots,Y_n$ i.i.d. with zero mean, $\mathrm{var}(Y_i)=s^2$, and $|Y_i|\le M$ a.s. [17]:

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right|\ge t\right)\le 2\exp\left(-\frac{t^2/2}{ns^2+Mt/3}\right).$$

Next, take $Y_i=X_i+X_{i+n}$. Then, $\mathrm{var}(Y_i)=2(\sigma^2-\tau)$ and $|Y_i|\le 2K$ a.s., thus

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{i=1}^{2n} X_i\right|\ge t\right)&=\mathbb{P}\left(\left|\sum_{i=1}^{n} Y_i\right|\ge t\right)\\&\le 2\exp\left(-\frac{t^2/2}{2n(\sigma^2-\tau)+2Kt/3}\right),\end{aligned}$$

yielding the desired inequality. Note that this is tighter than the usual Bernstein's inequality when t/n is dominated by σ^2/K as is the case in the sub-Gaussian regime. □

B. Proof of Proposition 10

Proof. Since Y is symmetric with respect to $-K/2$ and Z is symmetric with respect to $K/2$, the existence of such Y and Z is equivalent to

$$F(x)+F(-K-x)\le 1\le F(x)+F(K-x) \tag{3}$$

for all $|x|\le K$. Thus,

$$\begin{aligned}X+a(X)&=X+F^{-1}(1-F(X))\\&\overset{(a)}{\le} X+F^{-1}(F(K-x))\overset{\text{a.s.}}{=}K\end{aligned}$$

where (a) holds due to $1-F(x)\le F(K-x)$. Similarly, $X+a(X)\ge -K$ a.s. The two necessary conditions follow immediately from plugging $x=K/2$ or $x=-K/2$ in (3) (alternatively, observe $F_Y(-K/2)=F_Z(K/2)=1/2$).

The first sufficient condition implies that $F(x)\ge 1/2$ for all $x\ge 0$ and $F(x)\le 1/2$ for all $x\le 0$, which implies (3). The other two sufficient conditions both imply that $Y\preceq X\preceq Z$ with $Y\sim\mathsf{Unif}(-K,0)$ and $Z\sim\mathsf{Unif}(0,K)$. □

C. Proof of Proposition 13

We begin with a key lemma that analyzes the Lipschitz constant of g.

Lemma 16 (Lipschitz constant of antithetic sum)**.** *Let $f:\mathbb{R}^n\to\mathbb{R}$ be an L-Lipschitz (under ℓ^p-norm) and c-strongly isotonic function. Then, $g(\boldsymbol{x})=(f(\boldsymbol{x})+f(-\boldsymbol{x}))/2^{\frac{p-1}{p}}$ is $(L-c)$-Lipschitz continuous.*

Proof. Given some $\boldsymbol{x}=(x_1,\dots,x_n)$ and $\boldsymbol{y}=(y_1,\dots,y_n)$ with $\|\boldsymbol{x}-\boldsymbol{y}\|_p=d$, let $\boldsymbol{m}=(\max(x_1,y_1),\dots,\max(x_n,y_n))$ be the coordinate-wise maximum of $\boldsymbol{x}$ and $\boldsymbol{y}$. Then, $\boldsymbol{m}\ge\boldsymbol{x}$ and $\boldsymbol{m}\ge\boldsymbol{y}$. Next, let $\|\boldsymbol{m}-\boldsymbol{x}\|_p=a$ and $\|\boldsymbol{m}-\boldsymbol{y}\|_p=b$. Note that $\|(a,b)\|_p=d$. Then,

$$\begin{aligned}f(\boldsymbol{x})-f(\boldsymbol{y})&=f(\boldsymbol{x})-f(\boldsymbol{m})+f(\boldsymbol{m})-f(\boldsymbol{y})\\&\overset{(a)}{\le}-ca+f(\boldsymbol{m})-f(\boldsymbol{y})\overset{(b)}{\le}-ca+Lb\,,\end{aligned}$$

where (a) follows from $\boldsymbol{m}\ge\boldsymbol{x}$ and c-strong isotonicity, and (b) follows from L-Lipschitzness. Similarly, we can show that $f(\boldsymbol{x})-f(\boldsymbol{y})\ge -La+cb$ and $ca-LB\le f(-\boldsymbol{x})-f(-\boldsymbol{y})\le La-cb$. Thus,

$$|f(\boldsymbol{x})-f(\boldsymbol{y})+f(-\boldsymbol{x})-f(-\boldsymbol{y})|\le(L-c)(a+b), \tag{4}$$

which implies that

$$\begin{aligned}|g(\boldsymbol{x})-g(\boldsymbol{y})|&=\left|\frac{f(\boldsymbol{x})+f(-\boldsymbol{x})}{2^{\frac{p-1}{p}}}-\frac{f(\boldsymbol{y})+f(-\boldsymbol{y})}{2^{\frac{p-1}{p}}}\right|\\&=\left|\frac{f(\boldsymbol{x})-f(\boldsymbol{y})+f(-\boldsymbol{x})-f(-\boldsymbol{y})}{2^{\frac{p-1}{p}}}\right|\\&\overset{(a)}{\le}(L-c)\frac{a+b}{2^{\frac{p-1}{p}}}\overset{(b)}{\le}(L-c)\|\boldsymbol{x}-\boldsymbol{y}\|_p\,,\end{aligned}$$

where (a) follows from (4) and (b) is due to Hölder's inequality $|a+b|\le\|(a,b)\|_p\|(1,1)\|_{\frac{p}{p-1}}=2^{\frac{p-1}{p}}\|(a,b)\|_p$. □

With Lemma 16, proving Proposition 13 is straightforward.

Proof of Proposition 13. Per the standard Gaussian concentration inequality for Lipschitz functions [55], if $f:\mathbb{R}^n\to\mathbb{R}$ is L-Lipschitz and $\boldsymbol{X}\sim\mathcal{N}(0,1)^n$,

$$\mathbb{P}(|f(\boldsymbol{X})-\mathbb{E}[f(\boldsymbol{X})]|\ge t)\le 2\exp\left(-\frac{t^2}{2L^2}\right).$$

Applying this to g, which is $(L-c)$-Lipschitz under ℓ^2 norm due to Lemma 16, immediately yields the desired result. □

VII. Proofs of Stochastic Optimization

A. Proof of Theorem 14

This proof adapts standard ideas from, e.g., [60]–[62], which bound the error for a single iteration and then take the expectation with a telescoping sum. We begin by proving the following lemma.

Lemma 17 (Single iteration error bound)**.** *If $\eta\le\frac{1}{L}$, we have*

$$\|\nabla f(\boldsymbol{x}_t)\|_2^2\le\frac{2(\Delta_t-\mathbb{E}_{\boldsymbol{Z}_t}[\Delta_{t+1}])}{\eta}+\frac{L\eta}{2}(\sigma^2-c^2\tau dp),$$

for all $t\in\{1,\dots,T\}$, where $\Delta_t\triangleq f(\boldsymbol{x}_t)-f(\boldsymbol{x}^*)\ge 0$.

Proof. By smoothness we have

$$\begin{aligned}f(\boldsymbol{x}_{t+1})&\le f(\boldsymbol{x}_t)+\nabla f(\boldsymbol{x}_t)^{\mathrm{T}}(\boldsymbol{x}_{t+1}-\boldsymbol{x}_t)+\frac{L}{2}\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|_2^2\\&=f(\boldsymbol{x}_t)-\eta\nabla f(\boldsymbol{x}_t)^{\mathrm{T}}\boldsymbol{G}_t+\frac{L\eta^2}{2}\|\boldsymbol{G}_t\|_2^2,\end{aligned}$$

where $\boldsymbol{G}_t=(\boldsymbol{g}(\boldsymbol{x}_t,\boldsymbol{Z}_t)+\boldsymbol{g}(\boldsymbol{x}_t,\boldsymbol{a}(\boldsymbol{Z}_t)))/2$. Adding and subtracting $\nabla f(\boldsymbol{x}_t)$ from $\eta\nabla f(\boldsymbol{x}_t)^{\mathrm{T}}\boldsymbol{G}_t$, we have

$$\begin{aligned}&\eta\nabla f(\boldsymbol{x}_t)^{\mathrm{T}}(\boldsymbol{G}_t-\nabla f(\boldsymbol{x}_t)+\nabla f(\boldsymbol{x}_t))\\&=\eta\|\nabla f(\boldsymbol{x}_t)\|_2^2+\eta\nabla f(\boldsymbol{x}_t)^{\mathrm{T}}(\boldsymbol{G}_t-\nabla f(\boldsymbol{x}_t)).\end{aligned}$$

This result then yields

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) - \eta \|\nabla f(\boldsymbol{x}_t)\|_2^2 - \eta \nabla f(\boldsymbol{x}_t)^{\mathrm{T}}(\boldsymbol{G}_t - \nabla f(\boldsymbol{x}_t)) + \frac{L\eta^2}{2}\|\boldsymbol{G}_t - \nabla f(\boldsymbol{x}_t) + \nabla f(\boldsymbol{x}_t)\|_2^2.$$

Let us take the expectation $\mathbb{E}_{\boldsymbol{Z}_t}[\cdot]$ (with respect to $\boldsymbol{Z}_t$ conditioned on $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{t-1}$) on both sides to obtain

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{Z}_t}[f(\boldsymbol{x}_{t+1})] \\
&\leq f(\boldsymbol{x}_t) - \eta\|\nabla f(\boldsymbol{x}_t)\|_2^2 - \eta \nabla f(\boldsymbol{x}_t)^{\mathrm{T}} \mathbb{E}_{\boldsymbol{Z}_t}[\boldsymbol{G}_t - \nabla f(\boldsymbol{x}_t)] \\
&\qquad + \frac{L\eta^2}{2}\mathbb{E}_{\boldsymbol{Z}_t}\left[\|\boldsymbol{G}_t - \nabla f(\boldsymbol{x}_t) + \nabla f(\boldsymbol{x}_t)\|_2^2\right] \\
&\overset{(a)}{=} f(\boldsymbol{x}_t) - \eta\|\nabla f(\boldsymbol{x}_t)\|_2^2 \\
&\qquad + \frac{L\eta^2}{2}\|\nabla f(\boldsymbol{x}_t)\|_2^2 + \frac{L\eta^2}{2}\mathbb{E}_{\boldsymbol{Z}_t}\left[\|\boldsymbol{G}_t - \nabla f(\boldsymbol{x}_t)\|_2^2\right] \\
&\overset{(b)}{\leq} f(\boldsymbol{x}_t) - \eta\left(1 - \frac{L\eta}{2}\right)\|\nabla f(\boldsymbol{x}_t)\|_2^2 + \frac{L\eta^2}{2}.\frac{\sigma^2 - c^2\tau dp}{2},
\end{aligned}$$

where (a) follows from the unbiasedness of our oracle $\boldsymbol{g}$ (and by extension $\boldsymbol{G}_t$), and (b) follows from Theorem 6. Letting $\eta \leq \frac{1}{L}$ such that $1 - \frac{L\eta}{2} \geq \frac{1}{2}$, subtracting $f(\boldsymbol{x}^*)$ to both sides, and rearranging completes the proof. □

Next we, use Lemma 17 to show Theorem 14.

Proof of Theorem 14. From Lemma 17, we take the expectation with respect to $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_T$ to obtain

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_T}\left[\|\nabla f(\boldsymbol{x}_t)\|_2^2\right] \\
&\leq \frac{2\mathbb{E}_{\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_T}[\Delta_t - \Delta_{t+1}]}{\eta} + \frac{L\eta}{2}(\sigma^2 - c^2\tau dp),
\end{aligned}$$

using the fact that each $\boldsymbol{Z}_t$ is drawn i.i.d. from P_Z^d. Averaging over t, we obtain

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_T}\left[\|\nabla f(\boldsymbol{x}_t)\|_2^2\right] \\
&\leq \frac{2(\Delta_1 - \mathbb{E}_{\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_T}[\Delta_{t+1}])}{T\eta} + \frac{L\eta}{2}(\sigma^2 - c^2\tau dp).
\end{aligned}$$

Let $\hat{\boldsymbol{x}}$ be a random vector such that $\mathbb{P}(\hat{\boldsymbol{x}} = \boldsymbol{x}_t) = 1/T$. Then, by tower expectation,

$$\mathbb{E}_{\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_T,\hat{\boldsymbol{x}}}\left[\|\nabla f(\hat{\boldsymbol{x}})\|_2^2\right] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_T}\left[\|\nabla f(\boldsymbol{x}_t)\|_2^2\right].$$

Using the fact that $\mathbb{E}_{\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_T}[\Delta_{t+1}] \geq 0$ completes the proof. □

B. Proof of Theorem 15

Proof. This proof adapts standard ideas from, e.g., [63], [64]. For simplicity, let

$$\boldsymbol{G}(\boldsymbol{x}, \boldsymbol{Z}) = \frac{\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{Z}) + \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{a}(\boldsymbol{Z}))}{2}.$$

Note that $\mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{G}(\boldsymbol{x}, \boldsymbol{Z})] = \nabla f(\boldsymbol{x})$ and $\mathrm{var}_{\boldsymbol{Z}}[\boldsymbol{G}(\boldsymbol{x}, \boldsymbol{Z})] = (\sigma^2 - c^2\tau pd)/2$. From $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$ and L-smoothness, we have $\|\nabla f(\boldsymbol{x})\|_2^2 \leq L^2\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2$. Thus,

$$\mathbb{E}_{\boldsymbol{Z}}[\|\boldsymbol{G}(\boldsymbol{x}, \boldsymbol{Z})\|_2^2] \leq L^2\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 + \frac{\sigma^2 - c^2\tau pd}{2}. \tag{5}$$

In addition, from strong convexity, we have

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \geq \frac{\mu\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2}{2}. \tag{6}$$

Then, letting δ_t denote $\|\boldsymbol{x}_t - \boldsymbol{x}^*\|_2^2$ for simplicity, we have

$$\begin{aligned}
\delta_{t+1} &= \|\boldsymbol{x}_t - \boldsymbol{x}^* - \eta\boldsymbol{G}(\boldsymbol{x}_t, \boldsymbol{Z}_t)\|_2^2 \\
&= \delta_t - 2\eta(\boldsymbol{x}_t - \boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{G}(\boldsymbol{x}_t, \boldsymbol{Z}_t) + \eta^2\|\boldsymbol{G}(\boldsymbol{x}_t, \boldsymbol{Z}_t)\|_2^2.
\end{aligned}$$

Taking the expectation with respect to $\boldsymbol{Z}_t$, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{Z}_t}[\delta_{t+1}] \\
&= \delta_t - 2\eta(\boldsymbol{x}_t - \boldsymbol{x}^*)^{\mathrm{T}}\nabla f(\boldsymbol{x}_t) + \eta^2\mathbb{E}_{\boldsymbol{Z}_t}\left[\|\boldsymbol{G}(\boldsymbol{x}_t, \boldsymbol{Z}_t)\|_2^2\right] \\
&\overset{(a)}{\leq} (1 + L^2\eta^2)\delta_t - 2\eta(\boldsymbol{x}_t - \boldsymbol{x}^*)^{\mathrm{T}}\nabla f(\boldsymbol{x}_t) + \eta^2\frac{\sigma^2 - c^2\tau pd}{2} \\
&\overset{(b)}{\leq} (1 - \eta\mu + L^2\eta^2)\delta_t - 2\eta(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)) + \eta^2\frac{\sigma^2 - c^2\tau pd}{2} \\
&\overset{(c)}{\leq} (1 - 2\eta\mu + L^2\eta^2)\delta_t + \eta^2\frac{\sigma^2 - c^2\tau pd}{2},
\end{aligned} \tag{7}$$

where (a) follows from (5), (b) follows from strong convexity, and (c) follows from (6). Thus,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_t}[\delta_{t+1}] \\
&\overset{(a)}{\leq} (1 - 2\eta\mu + L^2\eta^2)^t\delta_1 \\
&\qquad + \eta^2\frac{\sigma^2 - c^2\tau pd}{2}\sum_{i=0}^{t-1}(1 - 2\eta\mu + L^2\mu^2)^i \\
&\overset{(b)}{\leq} (1 - 2\eta\mu + L^2\eta^2)^t\delta_1 + \eta\frac{\sigma^2 - c^2\tau pd}{4\mu - 2L^2\eta} \\
&\overset{(c)}{\leq} (1 - \eta\mu)^t\delta_1 + \eta\frac{\sigma^2 - c^2\tau pd}{2\mu},
\end{aligned}$$

where (a) follows from multiple applications of (7), and (b) and (c) follow from $\eta \leq \mu/L^2$, completing the proof. □

VIII. Conclusion

In this paper, we defined the antithetic index, explored its properties and connections to various concepts in optimal transport and signal processing, derived strong antithetic variance reduction inequalities, and demonstrated how they can be used to improve theoretical results on sampling methods. Note that our approach is not limited to the examples explored in this paper, and can be used to improve the variance dependence of any method that relies on sampling a known distribution. Moreover, although we only investigate the i.i.d. setting for simplicity when deriving multivariate inequalities such as Theorems 5 and 6, this restriction can be loosened by allowing variables to have different antithetic indices. When the underlying distribution is unknown, existing distribution estimation methods could potentially be used to estimate the CDF from the sample data so that antithetic variates can be generated. In addition, although we impose anti-Lipschitzness or strong isotonicity in order to guarantee quantitative gains in antithetic variance reduction, it is worth considering finer measures of the degree of non-linearity of quantile functions to loosen this restriction. We leave these directions of research to future work.

APPENDIX A
PROOF OF PROPOSITION 1

Proof. First, if $\boldsymbol{u}$ is positive, for all $\boldsymbol{x} \geq \boldsymbol{y}$ with $\boldsymbol{x} \neq \boldsymbol{y}$ we have $\boldsymbol{u}^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{y}) > 0$. Thus, if $g(\boldsymbol{x})\boldsymbol{u}$ is a monotone operator,

$$(g(\boldsymbol{x})\boldsymbol{u} - g(\boldsymbol{y})\boldsymbol{u})^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{y}) = (g(\boldsymbol{x}) - g(\boldsymbol{y}))\boldsymbol{u}^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{y}) \geq 0,$$

implying that $g(\boldsymbol{x}) - g(\boldsymbol{y}) \geq 0$, establishing that g is isotonic.

Next, we proceed to the strongly monotone case. Suppose there exists some positive vector $\boldsymbol{u} \in \mathbb{R}^d$ such that $g(\boldsymbol{x})\boldsymbol{u}$ is a c-strongly monotone operator. Then, we have by definition

$$(g(\boldsymbol{x})\boldsymbol{u} - g(\boldsymbol{y})\boldsymbol{u})^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{y}) \geq c\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \geq c\|\boldsymbol{x}-\boldsymbol{y}\|_\infty^2$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. If $\boldsymbol{x} > \boldsymbol{y}$, then

$$\begin{aligned}(g(\boldsymbol{x})\boldsymbol{u} - g(\boldsymbol{y})\boldsymbol{u})^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{y}) &= (g(\boldsymbol{x}) - g(\boldsymbol{y}))\boldsymbol{u}^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{y}) \\ &\leq (g(\boldsymbol{x}) - g(\boldsymbol{y}))\|\boldsymbol{u}\|_1\|\boldsymbol{x}-\boldsymbol{y}\|_\infty,\end{aligned}$$

and thus $(g(\boldsymbol{x}) - g(\boldsymbol{y}))\|\boldsymbol{u}\|_1\|\boldsymbol{x}-\boldsymbol{y}\|_\infty \geq c\|\boldsymbol{x}-\boldsymbol{y}\|_\infty^2$, which implies that $g(\boldsymbol{x}) - g(\boldsymbol{y}) \geq (c/\|\boldsymbol{u}\|_1)\|\boldsymbol{x}-\boldsymbol{y}\|_\infty$. Therefore, g is $c/\|\boldsymbol{u}\|_1$-strongly isotonic. □

APPENDIX B
STOCHASTIC DOMINATION PROOF OF ANTITHETIC SAMPLING

Chebyshev's association inequality is not the usual approach to derive the antithetic variates method for variance reduction [16] (also see [7], [22], [24]). Since proofs for the vanilla antithetic variates method can be difficult to find in the Monte Carlo literature, we provide a self-contained alternative proof for it that completes partial arguments presented in various formal and informal sources. We reiterate that although the ensuing result is known, the presentation below is complete and potentially more palatable to a statistical signal processing audience.

Proposition 18 (Antithetic Variates [7])**.** *Given a uniform random variable U on the interval $[0,1]$ and any monotone non-decreasing functions $f : [0,1] \to \mathbb{R}$ and $g : [0,1] \to \mathbb{R}$, we have $\operatorname{cov}(f(U), g(1-U)) \leq 0$.*

Proof. First, we may assume that f and g have codomain $\mathbb{R}_+$, i.e., $f : [0,1] \to \mathbb{R}_+$ and $g : [0,1] \to \mathbb{R}_+$, without loss of generality. Indeed, $f(0) = \min_{t\in[0,1]} f(t)$ and $g(0) = \min_{t\in[0,1]} g(t)$, and we can construct monotone non-decreasing functions $\tilde{f} : [0,1] \to \mathbb{R}_+$, $\tilde{f}(t) = f(t) + f(0)$ and $\tilde{g} : [0,1] \to \mathbb{R}_+$, $\tilde{g}(t) = g(t) + g(0)$ such that $\operatorname{cov}(\tilde{f}(U), \tilde{g}(1-U)) = \operatorname{cov}(f(U), g(1-U))$.

Define the PDF p on $[0,1]$ such that $p(t) = f(t)/\int_0^1 f(t)\,\mathrm{d}t$ (for $t \in [0,1]$), where the normalization constant must be finite as f is bounded. Let $P : [0,1] \to [0,1]$, $P(t) = \int_0^t p(u)\,\mathrm{d}u$ denote the corresponding CDF. Since p is monotone non-decreasing, P is convex. Hence, we have $P(t) = P((t)1 + (1-t)0) \leq tP(1) + (1-t)P(0) = t$ for all $t \in [0,1]$, because $P(0) = 0$ and $P(1) = 1$. This implies that the CDF P (first-order) stochastically dominates the CDF of U. Since $[0,1] \ni t \mapsto g(1-t)$ is monotone non-increasing, this dominance yields the following monotonicity relation between the expected values of $g(1-t)$ with respect to p and U:

$$\int_0^1 p(t)g(1-t)\,\mathrm{d}t \leq \int_0^1 g(1-t)\,\mathrm{d}t\,,$$

where the integrals are finite because f and g are bounded. Equivalently, we have

$$\begin{aligned}\mathbb{E}[f(U)g(1-U)] &= \int_0^1 f(t)g(1-t)\,\mathrm{d}t \\ &\leq \int_0^1 f(t)\,\mathrm{d}t \int_0^1 g(1-t)\,\mathrm{d}t = \mathbb{E}[f(U)]\,\mathbb{E}[g(1-U)]\,,\end{aligned}$$

i.e., $\operatorname{cov}(f(U), g(1-U)) \leq 0$, which completes the proof. □

REFERENCES

[1] A. Hashemi, D. Lee, and A. Makur, "Strong antithetic variance reduction inequalities," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Ann Arbor, MI, USA, June 22–27 2025, pp. 1–6.

[2] H. Kahn and A. W. Marshall, "Methods of reducing sample size in Monte Carlo computations," *Journal of the Operations Research Society of America*, vol. 1, no. 5, pp. 263–278, November 1953.

[3] J. Geweke, "Monte Carlo simulation and numerical integration," in *Handbook of Computational Economics*, H. M. Amman, D. A. Kendrick, and J. Rust, Eds. Amsterdam, Netherlands: Elsevier, 1996, vol. 1, pp. 731–800.

[4] M. Evans and T. Swartz, *Approximating integrals via Monte Carlo and deterministic methods*. Oxford, UK: Oxford University Press, 2000.

[5] S. Asmussen and P. W. Glynn, *Stochastic Simulation: Algorithms and Analysis*. New York, NY, USA: Springer, 2007.

[6] M. H. Kalos and P. A. Whitlock, *Monte Carlo Methods*. Hoboken, NJ, USA: John Wiley & Sons, 2008.

[7] R. Y. Rubinstein and D. P. Kroese, *Simulation and the Monte Carlo Method*, 3rd ed. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2017.

[8] A. Doucet and X. Wang, "Monte Carlo methods for signal processing: a review in the statistical signal processing context," *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 152–170, November 2005.

[9] C. A. Naesseth, F. Lindsten, and T. B. Schön, "High-dimensional filtering using nested sequential Monte Carlo," *IEEE Transactions on Signal Processing*, vol. 67, no. 16, pp. 4177–4188, July 2019.

[10] D. Luengo, L. Martino, M. Bugallo, V. Elvira, and S. Särkkä, "A survey of Monte Carlo methods for parameter estimation," *EURASIP Journal on Advances in Signal Processing*, vol. 2020, no. 1, p. 25, May 2020.

[11] S. Mohamed, M. Rosca, M. Figurnov, and A. Mnih, "Monte Carlo gradient estimation in machine learning," *Journal of Machine Learning Research*, vol. 21, no. 132, pp. 1–62, July 2020.

[12] V. Elvira and I. Santamaria, "Multiple importance sampling for symbol error rate estimation of maximum-likelihood detectors in MIMO channels," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1200–1212, February 2021.

[13] J. M. Hammersley and K. W. Morton, "A new Monte Carlo technique: Antithetic variates," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 52, no. 3, pp. 449–475, July 1956.

[14] E. C. Fieller and H. O. Hartley, "Sampling with control variables," *Biometrika*, vol. 41, no. 3/4, pp. 494–501, December 1954.

[15] G. S. Fishman and B. D. Huang, "Antithetic variates revisited," *Communications of the ACM*, vol. 26, no. 11, pp. 964–971, November 1983.

[16] S. Gal, R. Y. Rubinstein, and A. Ziv, "On the optimality and efficiency of common random numbers," *Mathematics and Computers in Simulation*, vol. 26, no. 6, pp. 502–512, December 1984.

[17] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford, UK: Oxford University Press, 2013.

[18] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Proceedings of the Advances in Neural Information Processing Systems 26 (NeurIPS)*, Lake Tahoe, NV, USA, December 5–10 2013, pp. 315–323.

[19] M. Wu, N. Goodman, and S. Ermon, "Differentiable antithetic sampling for variance reduction in stochastic variational inference," in *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, Naha, Japan, April 16–18 2019, pp. 2877–2886.

[20] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals and Systems*, vol. 2, no. 4, pp. 303–314, December 1989.

[21] A. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 930–945, May 1993.

[22] J. M. Hammersley and J. G. Mauldon, "General principles of antithetic variates," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 52, no. 3, pp. 476–481, July 1956.

[23] D. C. Handscomb and J. M. Hammersley, "Proof of the antithetic variates theorem for $n > 2$," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 54, no. 2, pp. 300–301, April 1958.

[24] W. Whitt, "Bivariate distributions with given marginals," *The Annals of Statistics*, vol. 4, no. 6, pp. 1280–1289, November 1976.

[25] C. M. Fortuin, P. W. Kasteleyn, and J. Ginibre, "Correlation inequalities on some partially ordered sets," *Communications in Mathematical Physics*, vol. 22, no. 2, pp. 89–103, June 1971.

[26] R. Ahlswede and D. E. Daykin, "An inequality for the weights of two families of sets, their unions and intersections," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 43, no. 3, pp. 183–185, September 1978.

[27] D. Achlioptas and K. Zampetakis, "A simpler proof of the four functions theorem and some new variants," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Espoo, Finland, June 26–July 1 2022, pp. 714–717.

[28] H. Ren, S. Zhao, and S. Ermon, "Adaptive antithetic sampling for variance reduction," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, Long Beach, CA, USA, June 9–15 2019, pp. 5420–5428.

[29] N. Le Roux, M. Schmidt, and F. Bach, "A stochastic gradient method with an exponential convergence rate for finite training sets," in *Proceedings of the Advances in Neural Information Processing Systems 25 (NeurIPS)*, Lake Tahoe, NV, USA, December 3–8 2012, pp. 2663–2671.

[30] M. Schmidt, N. Le Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," *Mathematical Programming*, vol. 162, no. 1, pp. 83–112, March 2017.

[31] A. Defazio, F. Bach, and S. Lacoste-Julien, "SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives," in *Proceedings of the Advances in Neural Information Processing Systems 27 (NeurIPS)*, Montreal, QC, Canada, December 8–13 2014, pp. 1646–1654.

[32] M. Mahdavi, L. Zhang, and R. Jin, "Mixed optimization for smooth functions," in *Proceedings of the Advances in Neural Information Processing Systems 26 (NeurIPS)*, Lake Tahoe, NV, USA, December 5–10 2013, pp. 674–682.

[33] R. Babanezhad Harikandeh, M. O. Ahmed, A. Virani, M. Schmidt, J. Konečný, and S. Sallinen, "Stopwasting my gradients: Practical SVRG," in *Proceedings of the Advances in Neural Information Processing Systems 28 (NeurIPS)*, Montreal, QC, Canada, December 7–12 2015, pp. 2251–2259.

[34] A. Jadbabaie, A. Makur, and D. Shah, "Gradient-based empirical risk minimization using local polynomial regression," *Stochastic Systems, INFORMS*, vol. 14, no. 4, pp. 363–402, December 2024.

[35] P. Zhao and T. Zhang, "Stochastic optimization with importance sampling for regularized loss minimization," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, Lille, France, July 6–11 2015, pp. 1–9.

[36] A. Katharopoulos and F. Fleuret, "Not all samples are created equal: Deep learning with importance sampling," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, Stockholm, Sweden, July 10–15 2018, pp. 2525–2534.

[37] M. Paul, S. Ganguli, and G. K. Dziugaite, "Deep learning on a data diet: Finding important examples early in training," in *Proceedings of the Advances in Neural Information Processing Systems 34 (NeurIPS)*, December 6–14 2021, pp. 20 596–20 607.

[38] R. M. Gower, M. Schmidt, F. Bach, and P. Richtárik, "Variance-reduced methods for machine learning," *Proceedings of the IEEE*, vol. 108, no. 11, pp. 1968–1983, November 2020.

[39] A. Rahimi and B. Recht, "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning," in *Proceedings of the Advances in Neural Information Processing Systems 21 (NeurIPS)*, Vancouver, BC, Canada, December 8–13 2008, pp. 1313–1320.

[40] B. Adcock, "Infinite-dimensional compressed sensing and function interpolation," *Foundations of Computational Mathematics*, vol. 18, no. 3, pp. 661–701, June 2018.

[41] H. Rauhut and R. Ward, "Sparse Legendre expansions via ℓ_1-minimization," *Journal of Approximation Theory*, vol. 164, no. 5, pp. 517–533, May 2012.

[42] P. Embrechts and M. Hofert, "A note on generalized inverses," *Mathematical Methods of Operations Research*, vol. 77, no. 3, pp. 423–432, June 2013.

[43] C. Sormani and C. Vega, "Null distance on a spacetime," *Classical and Quantum Gravity*, vol. 33, no. 8, pp. 1–29, March 2016.

[44] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Cham, Switzerland: Springer, 2011.

[45] B. F. Logan and L. A. Shepp, "Optimal reconstruction of a function from its projections," *Duke Mathematical Journal*, vol. 42, no. 4, pp. 645–659, December 1975.

[46] S.-L. Huang, A. Makur, G. W. Wornell, and L. Zheng, *Universal Features for High-Dimensional Learning and Inference*, ser. Foundations and Trends in Communications and Information Theory, A. Barg, Ed. Hanover, MA, USA: now Publishers Inc., February 2024, vol. 21, no. 1-2.

[47] L. Ambrosio, E. Brué, and D. Semola, *Lectures on optimal transport*, 2nd ed. Cham, Switzerland: Springer, 2021.

[48] E. C. Titchmarsh, *Introduction to the Theory of Fourier Integrals*, 2nd ed. Oxford, UK: Clarendon Press, 1948.

[49] K. Huang, Y. C. Eldar, and N. D. Sidiropoulos, "Phase retrieval from 1D Fourier measurements: Convexity, uniqueness, and algorithms," *IEEE Transactions on Signal Processing*, vol. 64, no. 23, pp. 6105–6117, December 2016.

[50] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, vol. 22, no. 3, pp. 400–407, September 1951.

[51] S. Bernstein, "Sur les fonctions absolument monotones," *Acta Mathematica*, vol. 52, no. 1, pp. 1–66, December 1929.

[52] D. V. Widder, *The Laplace Transform*. Princeton, NJ, USA: Princeton University Press, 1941.

[53] H. Taghavian, R. Drummond, and M. Johansson, "Logarithmically completely monotonic rational functions," *Automatica, Elsevier*, vol. 155, no. 111122, pp. 1–11, September 2023.

[54] J. Hadar and W. R. Russell, "Rules for ordering uncertain prospects," *The American economic review*, vol. 59, no. 1, pp. 25–34, March 1969.

[55] V. D. Milman and G. Schechtman, *Asymptotic Theory of Finite Dimensional Normed Spaces*. Berlin, Heidelberg: Springer, 1986.

[56] Y. Luo, X. Huo, and Y. Mei, "Implicit regularization properties of variance reduced stochastic mirror descent," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Espoo, Finland, June 26–July 1 2022, pp. 696–701.

[57] P. Jain and P. Kar, *Non-convex Optimization for Machine Learning*, ser. Foundations and Trends in Machine Learning, M. Jordan, Ed. Hanover, MA, USA: now Publishers Inc., 2017, vol. 10, no. 3–4.

[58] T. Popoviciu, "Sur les équations algébriques ayant toutes leurs racines réelles," *Mathematica*, vol. 9, pp. 129–145, 1935.

[59] V. M. Panaretos and Y. Zemel, "Statistical aspects of Wasserstein distances," *Annual review of statistics and its application*, vol. 6, no. 1, pp. 405–431, March 2019.

[60] S. Ghadimi and G. Lan, "Stochastic first-and zeroth-order methods for nonconvex stochastic programming," *SIAM journal on optimization*, vol. 23, no. 4, pp. 2341–2368, December 2013.

[61] R. Cosson, A. Jadbabaie, A. Makur, A. Reisizadeh, and D. Shah, "Low-rank gradient descent," *IEEE Open Journal of Control Systems*, vol. 2, pp. 380–395, October 2023.

[62] H. Li, J. Qian, Y. Tian, A. Rakhlin, and A. Jadbabaie, "Convex and non-convex optimization under generalized smoothness," in *Proceedings of the Advances in Neural Information Processing Systems 36 (NeurIPS)*, New Orleans, LA, USA, December 10–16 2023, pp. 40 238–40 271.

[63] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Review*, vol. 60, no. 2, pp. 223–311, January 2018.

[64] R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik, "SGD: General analysis and improved rates," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, Long Beach, CA, USA, June 9–15 2019, pp. 5200–5209.