THE STRONG, WEAK AND BENIGN GOODHART'S LAW: AN INDEPENDENCE-FREE FORMALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Goodhart's law is a famous adage in policy-making stating that "When a measure becomes a target, it ceases to be a good measure." As machine learning models and the optimization capacity to train them grow, increasing empirical evidence has supported this law, although it has not been formalized. Recently, a few attempts have been made to formalize Goodhart's law, either by categorizing variants of it, or by looking at how optimizing a *proxy metric* affects an *intended goal*. In this work, we relax the simplifying independence assumption of previous work, to study the effect of the coupling between the intended goal and the discrepancy on Goodhart's law. Our results show that when the tail of the goal strictly dominates the tail of the discrepancy, the dependence does not change the nature of Goodhart's law. However, in the light-tailed goal and heavy-tailed discrepancy case, we provide an example where the goal collapses (strong Goodhart), and the lighter the tail of the discrepancy, the faster the collapse. To the best of our knowledge, this is the first formalization of Goodhart's law that completely relaxes the independence assumption.

1 INTRODUCTION

From Charles Goodhart's remark in the context of monetary economics (Goodhart, 1975) to its reformulation by Keith Hoskin (1996) and its popularization by Marilyn Strathern (1997), Goodhart's law has remained mostly unformalized.

As metrics play an important role in designing artificial intelligence (AI) systems, much recent work on Goodhart's law has naturally come from AI research (Dotan et al., 2024; Birkstedt et al., 2023; Thomas & Uminsky, 2022; Hsia et al., 2023). In the context of training an AI system, a *proxy* metric (e.g., a loss function, a reward model or an expected gain) is set up to be optimized, and models are selected based on it. However, the metric being optimized is almost always just a proxy for the *goal*, an abstract objective capturing the most desirable properties for the AI model to have, properties that can hardly be encoded in a loss function (Amodei et al., 2016). In this work, we advance the mathematical formalization of Goodhart's law by studying how optimizing the proxy affects the goal.

When it comes to mathematically categorizing and formalizing Goodhart's law, three notable works took important steps[1] on which our work builds. First, Manheim & Garrabrant (2018) provided a categorization of Goodhart's law into distinct variants, without proving results about them or identifying the assumptions they require. Second, El-Mhamdi & Hoang (2024) used the covariance between the goal and the proxy and provided proofs for cases where Goodhart's law does not apply (optimizing the proxy is useful for the goal), cases where it applies in a *weak* form (optimizing the proxy does not benefit the goal, but does not harm it either) and cases where it applies in a *strong* form (optimizing the proxy harms the goal). Third, Kwa et al. (2024) argued for a notion of catastrophic Goodhart in the context of reinforcement learning from human feedback. Common to both El-Mhamdi & Hoang (2024) and Kwa et al. (2024) is an assumption of independence between the intended goal and the discrepancy. Our work removes this assumption and proposes a new categorization that subsumes those of El-Mhamdi & Hoang (2024) and Kwa et al. (2024).

[1] In addition, of course, to adjacent approaches, e.g., on reward hacking (Skalse et al., 2022), which are outside the scope of this study.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Throughout the paper, we model the setting with three real-valued random variables: a goal G (the real function we would want to optimize), a proxy metric M, and the discrepancy $\xi = M - G$ between them. Optimization is treated as a mechanism-agnostic selection procedure: rather than specifying a learning algorithm, we condition on the event $\{M > m\}$ and study the behavior of the goal G as the selection threshold m grows. Whether this selection benefits, stalls, or harms the goal is governed by the interplay between two features of the pair (G, ξ): the relative thickness of their distribution tails and their dependence structure. Previous formalizations assumed the discrepancy to be independent of the goal, an assumption that is hard to justify when the proxy is built from, or trained on, quantities correlated with the goal itself. Our contributions, which remove this assumption, are the following:

- We relax the independence assumption of previous mathematical formalizations of Goodhart's law, and derive general results that are agnostic to the dependence structure and rely on the assumption that the goal has a heavier tail than the discrepancy (either the goal is heavy-tailed and the discrepancy light-tailed, or both are heavy-tailed and their tail indices γ (goal) and α (discrepancy) satisfy $\frac{1}{\alpha} - \frac{1}{\gamma} > 0$).
- We conduct a detailed analysis of two cases. When the goal and the discrepancy are Gaussian, optimization succeeds while the metric becomes uninformative, which is an example of benign Goodhart. When the goal is exponential and the discrepancy is heavy-tailed, a specific dependence structure leads to an example of optimization failure (strong Goodhart), where the goal is minimized when the metric is maximized.
- We propose a taxonomy and formal characterization of Goodhart's law by introducing four regimes: *no*, *benign*, *weak* and *strong Goodhart*. Each characterizes a distinct outcome of optimization in our setting.
- We experimentally assess the robustness of our theoretical results in the Gaussian case by sampling Gaussian marginals coupled through non-Gaussian copulas, and find results consistent with our theory. We also experimentally test cases where our general results do not apply (when the tail of the discrepancy is thicker than the tail of the goal), which mostly lead to weak Goodhart.

Paper structure. The rest of this paper is organized as follows. Section 2 discusses related work. Section 3 introduces the model used for our analysis; in Section 4 we first provide a comprehensive overview of our results, followed by definitions and formal statements (detailed proofs are available in the Appendix). Section 5 provides an experimental assessment of our theoretical contributions. Finally, Section 6 discusses our results and situates them within both ongoing efforts and possible future work.

2 Related work

Several branches of the literature have separately considered Goodhart's law, with diverse approaches and formalizations.

General work on Goodhart's law. In decision analysis, Smith & Winkler (2006) identified the "optimizer's curse": selecting the alternative with the best noisy estimate systematically inflates the estimate relative to the realized value, an early expression of Goodhart's mechanism through selection. Manheim (2023) and Manheim & Garrabrant (2018) provide insights into metric design, offering, respectively, an overview of the potential flaws of metrics and a taxonomy of Goodhart cases, while not fully formalizing the problem. Zhuang & Hadfield-Menell (2020) devise a general framework for AI over-optimization and derive results in the case of constrained resources and partially specified goals; their setup is inspired by incomplete contracting, but its reliance on state-space descriptions makes it more suitable for reinforcement learning.

Reinforcement learning. Reinforcement learning (RL) has naturally been a favored setting for formalizing Goodhart's law, since the most notable cases of reward hacking in AI have appeared in RL settings (Amodei et al., 2016; Gao et al., 2022; Pan et al., 2022). An early formalization is Everitt et al. (2017), which models reward corruption as a generalized Markov decision process (MDP) and shows that no agent can overcome arbitrary corruption without additional assumptions. Pan et al. (2022) provide one of the first systematic empirical studies of reward hacking, showing that more capable agents exploit misspecified rewards more aggressively, with phase transitions at which the true reward sharply drops. The first part of Kwa et al. (2024) shows that Kullback–Leibler (KL) regularization is ineffective at preventing reward hacking: with a heavy-tailed reward error, a

policy with an arbitrarily high proxy reward but low KL penalty and low true reward always exists. Karwowski et al. (2024) provide a geometric explanation of Goodhart's law in RL and devise a provably Goodhart-resilient early-stopping criterion. Skalse et al. (2022) introduce a formalization of reward hacking in policy learning and prove several results on the existence of hackable pairs of reward functions over different policy sets. Laidlaw et al. (2025) characterize reward hacking through the correlation between the true and proxy rewards on states visited by a reference policy π_{ref}, and show that this correlation predicts hacking empirically.

Supervised learning. In the context of supervised learning, Hennessy & Goodhart (2020) provide microfoundations for machine learning (ML) models in order to make them robust to the Lucas critique of classical Keynesian models (Lucas, 1976): a regulator makes predictions in a setting where, at test time, covariates can be manipulated by an agent seeking a more favorable decision. This gaming of a fixed decision rule is studied in strategic classification (Hardt et al., 2015), generalized by performative prediction (Perdomo et al., 2020), where the deployed model itself shifts the data distribution.

Optimization as conditioning. Kwa et al. (2024) and El-Mhamdi & Hoang (2024) provide a general formalization of Goodhart's law using an optimizer model based on randomly selecting a model among the best-performing ones on the metric M. They show that whether Goodhart's law holds depends on the relative tail thickness of the goal and the discrepancy; however, this was done under the assumption that the discrepancy is independent of the goal. Our work follows the same approach but removes the independence assumption.

3 MODEL

Notation. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, where Ω is a general sample space. Elements of Ω are written as ω. We consider three measurable functions $G : \Omega \to \mathbb{R}$, $M : \Omega \to \mathbb{R}$ and $\xi : \Omega \to \mathbb{R}$, representing respectively the goal (the function we would want to optimize), the proxy metric being optimized (a surrogate for G), and the discrepancy between them. They satisfy the relation $M(\omega) = G(\omega) + \xi(\omega)$. As M, G and ξ are measurable, they induce random variables on Ω. Following standard convention, the dependence on ω will be omitted when clear from context. For any random variable X on Ω, we denote by $\mathfrak{S}_X$ its support and $\mathfrak{S}_X^O$ its interior. We denote by $\overline{\mathfrak{S}}_X = \sup \mathfrak{S}_X$ and $\underline{\mathfrak{S}}_X = \inf \mathfrak{S}_X$ the upper and lower endpoints of the support, respectively.

Modeling. We model the optimization of the proxy metric M as a selection procedure. This consists in conditioning on the event $\{M > m\}$ and taking the limit $m \to \overline{\mathfrak{S}_M}$, which corresponds to selecting models that exceed increasingly high thresholds on the proxy metric.

Example 1. *Consider training a neural network with parameters $\theta \in \mathbb{R}^d$ on data $(X_i, Y_i)_{i=1}^n \in (\mathbb{R}^l)^n$. Define the sample space as $\Omega = \mathbb{R}^d \times \mathbb{R}^{ln}$ equipped with the product σ-algebra $\mathcal{B}(\mathbb{R}^d) \otimes \mathcal{B}(\mathbb{R}^{ln})$ and the product Lebesgue measure $\lambda^d \otimes \lambda^{ln}$. A sample $\omega = (\omega_1, \omega_2) \in \Omega$ consists of model parameters $\omega_1 \in \mathbb{R}^d$ and data $\omega_2 \in \mathbb{R}^{ln}$. We define three measurable functions: $M(\omega)$ (training reward), $G(\omega)$ (generalization reward), and $\xi(\omega) = M(\omega) - G(\omega)$ (the discrepancy).*

Taxonomy. To capture the different cases that might arise from optimization, we summarize in Table 1 a taxonomy of four outcomes. This taxonomy covers most situations and allows us to categorize our results.

- No Goodhart covers situations where Goodhart's law does not apply: optimizing the proxy M also optimizes the goal G and the proxy stays informative in the limit.
- Benign Goodhart captures cases in which, when the proxy M is optimized, the goal G is also maximized, but M becomes uninformative about G in the limit.
- Weak Goodhart covers situations where optimization of the proxy is inefficient or counterproductive, in the sense that it does not increase the value of the intended goal G or even decreases it. In both cases, the expected value of the goal G remains bounded away from its supremum after optimization.
- Strong Goodhart is a genuine failure mode of proxy optimization. In this case, when the metric M is maximized, the goal G is actually minimized: as M goes to its supremum, G goes to its infimum.

	Qualitative definition	Formal definition
No Goodhart	During optimization, the proxy stays informative and the goal G goes to its maximum value	$\exists\, l \in (0,1)$, $\mathrm{Corr}\,(G, M \mid M > m) \underset{m \to \overline{\mathfrak{S}_M}}{>} l$ and $\mathbb{E}\,[G \mid M > m] \underset{m \to \overline{\mathfrak{S}_M}}{\longrightarrow} \overline{\mathfrak{S}_G}$
Benign	Despite the proxy's decreasing informativeness, the goal is maximized	$\mathrm{Corr}\,(G, M \mid M > m) \underset{m \to \overline{\mathfrak{S}_M}}{\longrightarrow} 0$ and $\mathbb{E}\,[G \mid M > m] \underset{m \to \overline{\mathfrak{S}_M}}{\longrightarrow} \overline{\mathfrak{S}_G}$
Weak	The expected value of the goal is bounded away from its maximum value during optimization	$\exists\, \delta, \varepsilon \in \mathfrak{S}_G^O, \exists m_0 \in \mathfrak{S}_M, \forall m > m_0,$ $\delta < \mathbb{E}\,[G \mid M > m] < \varepsilon$
Strong	The goal goes to its minimum value during the optimization of the proxy	$\mathbb{E}\,[G \mid M > m] \underset{m \to \overline{\mathfrak{S}_M}}{\longrightarrow} \underline{\mathfrak{S}_G}$

Table 1: Qualitative and formal definitions of the four Goodhart outcomes.

4 RESULTS

Our results fall into two groups.

General results. Theorems 1 and 2 show that weak and strong Goodhart cannot occur when the goal has a thicker tail than the discrepancy:

- When the goal G is heavy-tailed and the discrepancy ξ is light-tailed (Theorem 1), the conditional expectation $\mathbb{E}[G|M > m]$ is bounded below by $m + o(m)$, which rules out weak and strong Goodhart.
- When the goal and the discrepancy are both heavy-tailed, with tail indices γ and α respectively, if $1/\alpha - 1/\gamma > 0$, the same lower bound holds (Theorem 2).

These results hold without any assumption on the dependence structure. However, they leave open the question of when benign Goodhart occurs with this tail structure.

Two explicit couplings. We complement these results with two tractable couplings (Section 4.4). When (G, ξ) is a Gaussian vector with $\mathrm{Cov}(G, M) > 0$ (e.g., $\mathrm{Var}(G) > \mathrm{Var}(\xi)$), $\mathbb{E}[G \mid M > m]$ grows linearly in m while $\mathrm{Corr}(G, M \mid M > m) \to 0$: benign Goodhart. When G is exponential and ξ is Pareto, with a coupling that reserves large discrepancies for small goals, $\mathbb{E}[G \mid M > m] \sim (\eta/m)^{b-1} \to 0$: strong Goodhart, whereas independence would only yield weak Goodhart; and the lighter the tail of the discrepancy, the faster the collapse.

4.1 DEFINITIONS

We need the following definitions.

Definition 1 (Slowly varying function)**.** *A function f is slowly varying if it is positive, measurable and satisfies $\lim_{t\to\infty} \frac{f(at)}{f(t)} = 1$ for every $a > 0$.*

It follows from the definition that a slowly varying function grows or decays more slowly than any power function. In the following definitions, l denotes a slowly varying function and F_X denotes the cumulative distribution function of a random variable X:

Definition 2 (Heavy-tailed random variables)**.** *We say that a random variable X is right heavy-tailed of index γ if $F_X(t) \underset{t\to+\infty}{=} 1 - \frac{l(t)}{t^{1/\gamma}}$.*

We say that a random variable X is left heavy-tailed of index γ if $F_X(t) \underset{t\to-\infty}{=} \frac{l(|t|)}{|t|^{1/\gamma}}$.

We define heavy-tailedness by requiring the survival function to be equal to $\frac{l(x)}{x^{1/\gamma}}$ at infinity. This definition is a compromise between generality and tractability. More general classes exist, such as the subexponential class; however, they do not yield general tail asymptotics. Moreover, under mild additional assumptions, heavy-tailedness as we define it describes the stationary distribution of

	G \ ξ	Light tail	Heavy tail
Assuming independence (1) (El-Mhamdi & Hoang, 2024) (2) (Kwa et al., 2024)	Light tail	(1,2) Benign Goodhart or no Goodhart	(1,2) Weak Goodhart worsening with tail **thickness** of the discrepancy
	Heavy tail	(1) No Goodhart with a Pareto-distributed G and various light-tailed or bounded distributions for ξ	(1) With Pareto distributions on G and ξ, the relevance of the proxy depends on the relative tail shapes of G and ξ
No assumption on independence (**This paper**)	Light tail	**Benign or strong Goodhart** in the Gaussian case	**Strong Goodhart** for a specific dependence between an exponential goal and a Pareto discrepancy, worsening as the discrepancy's tail gets lighter
	Heavy tail	**No** or **benign Goodhart**	**No** or **benign Goodhart** when the tail of the goal is thicker than the tail of the discrepancy

Table 2: Summary of results without assumptions on the learning paradigm. Kwa et al. (2024) also obtain results without independence, but specifically for reinforcement learning from human feedback (RLHF). Note that our definition of weak Goodhart corresponds to the definition of catastrophic Goodhart in Kwa et al. (2024), and that the definition of weak Goodhart in El-Mhamdi & Hoang (2024) may correspond to either our weak or our benign Goodhart definition.

stochastic recurrence equations such as $X_n = A_n X_{n-1} + B_n$, where A_n and B_n are random variables with finite moments (Kesten, 1973). These properties make this class well suited to modeling systems in which feedback loops may occur.

Definition 3 (Light-tailed random variables)**.** *We say that a random variable X is right light-tailed if $\frac{1-F_X(t)}{e^{-st}} \underset{t\to\infty}{\to} 0$ for some $s > 0$.*

We say that a random variable X is left light-tailed if $\frac{F_X(t)}{e^{st}} \underset{t\to-\infty}{\to} 0$ for some $s > 0$.

A random variable that is both left and right light-tailed has finite moments of all orders and concentrates around $\mathbb{E}[X]$, which makes large deviations rare. In particular, any sub-exponential random variable (in the concentration sense, i.e., with finite exponential moments) is both right and left light-tailed.

Definitions 4 and 5 introduce the two models solved explicitly in Section 4.4. They describe specific dependence structures and are not meant to be general: both derivations only illustrate what can happen under certain dependence structures.

Definition 4 (Gaussian vector)**.** *We say that a random vector (X, Y) forms a Gaussian vector with covariance matrix Σ, and write $(X, Y) \sim \mathcal{N}(0, \Sigma)$, if for $\Sigma = \begin{pmatrix} a & c \\ c & b \end{pmatrix}$ with $ab > c^2$, we have $\mathrm{Var}(X) = a$, $\mathrm{Var}(Y) = b$ and $\mathrm{Cov}(X, Y) = c$ with the density of the vector being $\frac{1}{2\pi\sqrt{\det(\Sigma)}} \exp\!\big(-\frac{1}{2} x^\top \Sigma^{-1} x\big)$.*

Definition 5 (Exponential goal, Pareto discrepancy)**.** *Let $X \sim \mathcal{E}(1)$ and, for $\eta > 0$ and $b > 2$, let Y be given conditionally on X by*

$$\mathbb{P}(Y > y \mid X) = \exp\Big(- X\big[(y/\eta)^{b-1} - 1\big]\Big), \qquad y \geq \eta.$$

Here, conditionally on X, $(Y/\eta)^{b-1} - 1$ is exponential with rate X. Integrating out X gives $\mathbb{P}(Y > y) = (\eta/y)^{b-1}$: marginally, Y is Pareto with scale η and density proportional to y^{-b}, which we write $Y \sim \mathcal{P}(\eta, b)$, while the coupling reserves large discrepancies for small values of the goal.

4.2 Heavy-tailed goal and light-tailed discrepancy

Here, we consider the general case where the goal is heavy-tailed and the discrepancy is light-tailed, without specifying the dependence structure or restricting to particular distributions.

In this case, we derive a lower bound on the expected value of the goal G after optimization, as shown in the following theorem.

Theorem 1. *Let G be heavy-tailed with tail index γ, where $1/\gamma > 1$, and slowly varying function l, with $\mathbb{E}\left[|G|^p \mathbb{1}\{G < 0\}\right] < \infty$ for some $p > 1$. Let ξ be right and left light-tailed, and set $M := G + \xi$. Then*

$$\mathbb{E}\left[G | M > m\right] \underset{m \to \infty}{\geq} m + o(m).$$

Theorem 1 shows that weak and strong Goodhart are impossible in this case. Indeed, the lower bound shows that $\mathbb{E}\left[G|M > m\right]$ goes to infinity. This result is independent of the dependence structure between the goal G and the discrepancy ξ, since the relative tail thickness counteracts any unfavorable dependence structure. This is why we need both the left and right tails of ξ to decay exponentially for this result to hold. Otherwise, one can construct a dependence structure such that the left tail of the discrepancy (where it is negative) compensates for the large values of the goal G. This appears in the proof through the choice of t_m, which balances the coverage of the support of ξ against the decay of $\mathbb{P}\left(G > m + t_m\right)$. A left tail that is too heavy makes this balance impossible.

4.3 Heavy-tailed goal and heavy-tailed discrepancy

We now consider the case where both the goal and the discrepancy are heavy-tailed, with a tail imbalance favorable to the goal ($1/\alpha - 1/\gamma > 0$, where α and γ are the tail indices of ξ and G, respectively), without specifying the dependence structure or restricting to particular distributions.

Theorem 2. *Suppose that G is heavy-tailed with tail index γ, and that the left tail of G is no heavier than its right tail, in the sense that $\mathbb{E}\left[|G|^{1/\gamma} \mathbb{1}\{G < 0\}\right] < \infty$. Suppose that the left and right tails of ξ are at most as heavy as those of a heavy-tailed variable of index α, and define $M := G + \xi$. Assume further that $1/\alpha - 1/\gamma > 0$ and $1/\gamma > 1$. Then*

$$\mathbb{E}\left[G | M > m\right] \underset{m \to \infty}{\geq} m + o(m).$$

The detailed proof is provided in Section B.2.

The lower bound of Theorem 2 shows that strong and weak Goodhart are impossible in this case, since $\mathbb{E}\left[G|M > m\right]$ necessarily goes to infinity. This result is also independent of the dependence structure: as in the light-tailed case, the condition $1/\alpha - 1/\gamma > 0$ counteracts any unfavorable dependence structure. Without this assumption, as in the heavy-tailed goal and light-tailed discrepancy case, the proof breaks down. Indeed, the balancing term would then be $t_m \underset{m \to +\infty}{\sim} m^\kappa$ with $\kappa > \alpha/\gamma \geq 1$. Balancing the coverage of the support of ξ against the decay of $\mathbb{P}\left(G > m + t_m\right)$ is then impossible without further assumptions.

Theorems 1 and 2 rule out weak and strong Goodhart; however, they say nothing about how informative the proxy remains about the goal G. Whether benign or no Goodhart occurs therefore remains open[2].

4.4 Two explicit couplings: dependence decides

When the tail of the goal does not dominate that of the discrepancy, Theorems 1 and 2 do not apply and the dependence structure becomes decisive. We illustrate this with two couplings for which

[2]Note that to correctly define the correlation we need the variance to be finite, which constrains the goal and discrepancy to be in L^2.

all quantities of interest can be computed explicitly: a Gaussian one, where optimization succeeds while the proxy loses its informativeness, and an exponential–Pareto one, where the goal is driven to its minimum.

Proposition 3. *Let* $M = G + \xi$.

(i) Gaussian coupling. *Suppose that* $(G, \xi) \sim \mathcal{N}(0, \Sigma)$ *as defined in Definition 4. If* $a + c > 0$*, which holds in particular whenever* $\mathrm{Var}(G) > \mathrm{Var}(\xi)$*,*

$$\mathbb{E}[G \mid M > m] \underset{m\to\infty}{\sim} \frac{a+c}{a+b+2c}\, m,$$
$$\mathrm{Corr}(G, M \mid M > m) \underset{m\to\infty}{\sim} \frac{(a+c)\sqrt{a+b+2c}}{\sqrt{ab-c^2}}\, \frac{1}{m}.$$

(ii) Exponential–Pareto coupling. *Suppose that* $G \sim \mathcal{E}(1)$ *and* $\xi \sim \mathcal{P}(\eta, b)$*, with the dependence structure of Definition 5. Then*

$$\mathbb{E}[G \mid M > m] \underset{m\to\infty}{\sim} \left(\frac{\eta}{m}\right)^{b-1},$$
$$\mathbb{E}[\xi \mid M > m] \underset{m\to\infty}{\sim} \frac{b-1}{b-2}\, m.$$

Higher-order expansions and proofs are given in Section A.1 for (i) and Sections A.2 to A.5 for (ii); the computations for (ii) have been verified symbolically with SymPy (Meurer et al., 2017) in the supplementary material.

Benign, weak or strong Goodhart in the Gaussian case. Since (G, M) is Gaussian, $G = \beta M + \varepsilon$ with $\beta = \frac{a+c}{a+b+2c}$, where ε is independent of M and has variance $\frac{ab-c^2}{a+b+2c}$. Selection thus acts on G only through M: $\mathbb{E}[G \mid M > m] = \beta\, \mathbb{E}[M \mid M > m]$ and $\mathbb{E}[M \mid M > m] \sim m$. Meanwhile, $\mathrm{Var}(M \mid M > m)$ decays like m^{-2}, whereas the variance of ε is unaffected by the selection, so the correlation vanishes. The regime is therefore determined by the sign of β, that is, by the sign of $a + c$, since $a + b + 2c = \mathrm{Var}(M) > 0$. If $a + c > 0$, then $\mathbb{E}[G \mid M > m] \to +\infty$: this is benign Goodhart. The condition $a + c > 0$ holds in particular whenever $\mathrm{Var}(G) > \mathrm{Var}(\xi)$, since positive definiteness of the covariance matrix yields $|c| < \sqrt{ab} < a$. Conversely, if $a + c < 0$, then $\mathbb{E}[G \mid M > m] \to -\infty$: this is strong Goodhart. Finally, when $a + c = 0$, the jointly Gaussian variables G and M are uncorrelated, hence independent, so that $\mathbb{E}[G \mid M > m] = \mathbb{E}[G] = 0$ for every m: selection on M has no effect on the goal, which is weak Goodhart.

Strong Goodhart in the exponential goal and Pareto discrepancy case. In the setting of Definition 5, given $G = g$, the discrepancy exceeds a level m with probability $\exp(-g[(m/\eta)^{b-1} - 1])$, which is non-negligible only if $g \lesssim (\eta/m)^{b-1}$. As the light-tailed goal cannot reach high levels by itself, a high threshold on M is exceeded almost exclusively through a large discrepancy, hence through a small goal: $\mathbb{E}[G \mid M > m] \to 0 = \inf S_G$, while ξ carries the whole excess ($\mathbb{E}[\xi \mid M > m]$ is asymptotically the conditional mean $\mathbb{E}[\xi \mid \xi > m]$ of its Pareto marginal). This is strong Goodhart, whereas the same marginals under independence only yield weak Goodhart, with $\mathbb{E}[G \mid M > m] \to \mathbb{E}[G]$ (El-Mhamdi & Hoang, 2024; Kwa et al., 2024): the dependence turns weak Goodhart into strong Goodhart. The rate is counterintuitive: the goal collapses (strong Goodhart), and the lighter the tail of the discrepancy, the faster the collapse, since a lighter tail requires an even smaller goal for the discrepancy to exceed m.

Remark 1. *Note that in the Gaussian case,* $a + c = \mathrm{Cov}(G, M)$ *controls the regime, and strong Goodhart arises only when the proxy is negatively correlated with the goal before any optimization. The exponential–Pareto coupling needs no such thing: with* $\eta = 1$*, the proxy is highly correlated with the goal (*$\mathrm{Corr}(G, M) \approx 0.88$ *for* $b = 5$ *and* 0.99 *for* $b = 10$*), and yet* $\mathbb{E}[G \mid M > m] \to 0$*.*

5 Experiments

We conduct experiments to assess the robustness of our findings in the Gaussian case and to show the variety of behaviors that can arise with a heavy-tailed goal and a heavy-tailed discrepancy with unfavorable tail balance. We use the Python package `clayton` (Boulin, 2023) to sample from copulas. The figures include error bands showing the 5th and 95th percentiles.

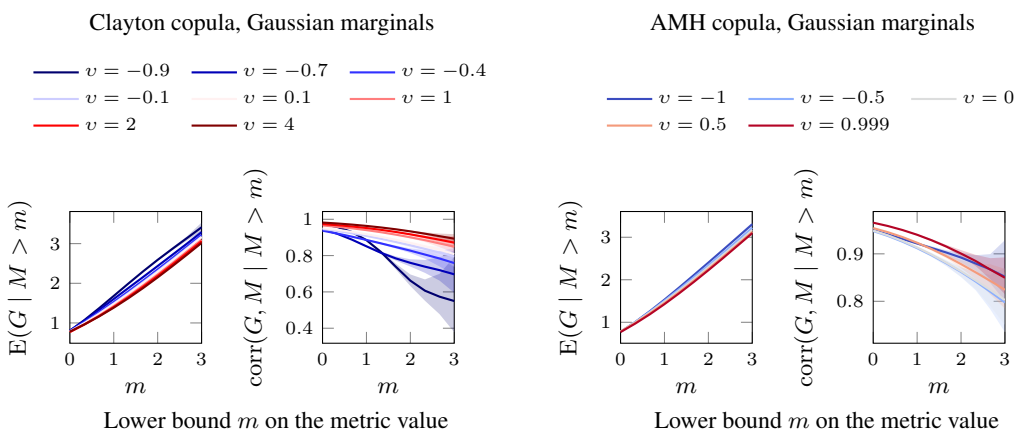


Figure 1: Gaussian marginals with different copulas, i.e., dependence structures. Colors range from deep red (strong positive dependence) to deep blue (strong negative dependence).

5.1 Experimental design

In each experiment, we draw 100,000 samples of the goal G and of the discrepancy ξ. The pair (G, ξ) is drawn from specific copulas (i.e., dependence structures) and marginals.

In Fig. 1, (G, ξ) is sampled with the Clayton copula (Clayton, 1978, left panels) and the Ali–Mikhail–Haq (AMH) copula (Ali et al., 1978, right panels) with Gaussian marginals of variance 1 for G and 0.2 for ξ.

In Fig. 2, (G, ξ) is sampled with the Clayton copula (left) and with a Gaussian copula (right). G has a heavy-tailed distribution of tail index $1/5$ (its survival function decays as $1/x^5$), while ξ has a heavy-tailed distribution of tail index $1/2$ (its survival function decays as $1/x^2$).

The copula parameters υ range from negative values (close to countermonotonicity) to positive values (close to comonotonicity). Each copula encodes a different dependence structure. In the experiments with the Clayton copula, the parameter ranges from $\upsilon = -0.9$ to $\upsilon = 4$, while in the experiments with the AMH copula, it ranges from $\upsilon = -1$ to $\upsilon = 0.999$. For the Gaussian copula, the correlation parameter ranges from $\upsilon = -0.61$ to $\upsilon = 0.61$.

After drawing (G, ξ), we compute $M = G + \xi$ and keep the pairs (G, M) for which M exceeds a threshold m. From these pairs, we estimate $\mathbb{E}[G|M > m]$ and $\mathrm{Corr}(G, M \mid M > m)$. In each plot, the x-axis shows the threshold m, and the y-axis shows either the conditional expectation or the correlation.

5.2 Results

Figure 1. With both Clayton and AMH copulas, we see that the conditional expectation of the goal increases as the dependence parameter becomes more negative. For the Clayton copula, the conditional expectation of the goal is 0.6 higher for $\upsilon = -0.9$ than for $\upsilon = 4$. The same pattern holds for the AMH copula, although the difference is smaller. The correlation decreases to 0 in both cases, faster under negative dependence.

Figure 2. For the Clayton copula (left plot) under positive dependence ($\upsilon > 0$), the conditional expectation of the goal first rises to 1.75 and then stabilizes about 0.7 above its mean. However, under negative dependence ($\upsilon < 0$), the conditional expectation of the goal decreases to near its minimum value (1). For the Gaussian copula under positive dependence ($\upsilon > 0$), the conditional expectation of the goal first grows to 3 and then seems to stabilize, except for the strongest dependence ($\upsilon = 0.61$). Under negative dependence ($\upsilon < 0$), the conditional expectation of the goal first increases, then decreases to near the unconditional mean $\mathbb{E}[G]$.

Discussion of the results. Figure 1 suggests that the pattern derived for Gaussian vectors also holds for other copulas.

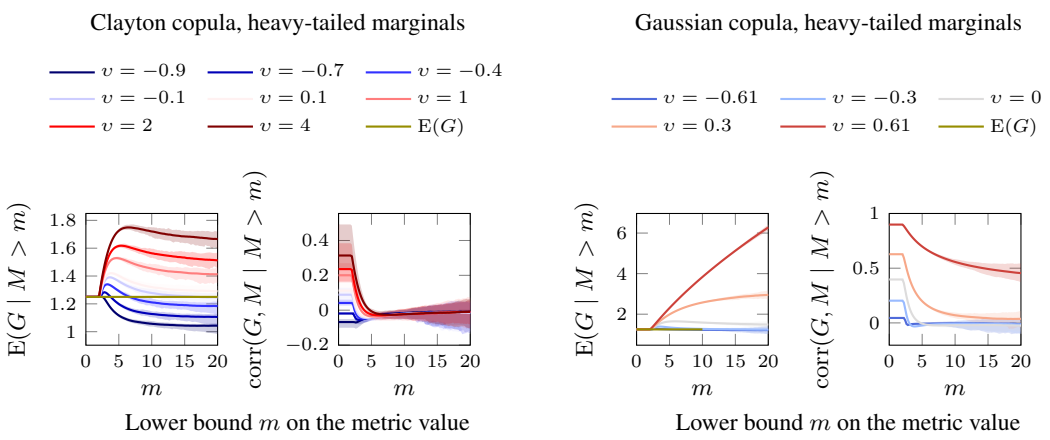


Figure 2: Pareto(1/5) goal and Pareto(1/2) discrepancy with different dependence structures. Colors range from deep red (strong positive dependence) to deep blue (strong negative dependence).

From Fig. 2, we draw two conclusions:

- When the dependence is favorable to the goal (i.e., G and ξ are positively dependent), the conditional mean of the goal appears to stabilize above its unconditional mean, which suggests that optimization only helps up to a point (and argues for early stopping: selecting instances where the metric is above 20 brings almost no further benefit to the goal).
- Under unfavorable dependence, optimization appears counterproductive: the conditional mean of the goal falls below its unconditional mean.

6 Discussion

Whether Goodhart's law occurs has recently been shown to depend on the tail thickness of both the intended goal and the discrepancy between that goal and the proxy metric. Our results reveal a more nuanced interplay between tail thickness and dependence structure.

We prove that weak and strong Goodhart cannot occur in two cases: a heavy-tailed goal with a light-tailed discrepancy, and a heavy-tailed goal with a heavy-tailed discrepancy whose tail imbalance favors the goal. However, the two other cases we analyze, namely the Gaussian case and the exponential goal with a heavy-tailed discrepancy, show the importance of the dependence structure when the relative tail thickness is not favorable to the goal or when both tails are similar. Indeed, with an exponential goal and a heavy-tailed discrepancy, the dependence structure turns weak Goodhart (under independence) into strong Goodhart. Moreover, this strong Goodhart effect worsens as the tail of the discrepancy gets lighter. Together with our previous results, this suggests that tail thickness does not *trigger* Goodhart's law but rather *enables* it. When the tails of the goal and the discrepancy are comparable, or unbalanced in favor of the discrepancy, some dependence structures can trigger Goodhart's law under optimization.

Theoretical guarantees on the possibility of harmful AI behavior before or at test time are key to mitigating global risks from AI. Moreover, there have been repeated calls for probabilistic formulations of such guarantees to inform AI safety research (Bengio et al., 2025). By providing such probabilistic formulations, our paper highlights two decisive factors for misalignment: heavy tails enable it, while the dependence structure actually triggers it. Overall, this work contributes to a better theoretical understanding of AI misalignment, and can serve as the basis for a Goodhart-centered research agenda on AI alignment.

AI USE STATEMENT

In this work, we used generative AI tools to polish text and mathematical phrasing, to implement methods, and to assist with translation. We did not use generative AI tools to assist in writing proofs, propose methods, proofs or hypotheses, generate synthetic data, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, design or provide feedback on research methodology or experiments, support qualitative and thematic data analysis, interpret results, or clean and reformat datasets. The remaining categories do not apply to this work. Additionally, we used generative AI tools to create or modify scientific figures or images, summarize or analyze existing literature, discover research topics or identify gaps, edit this paper to improve readability, and propose a title or keywords for this paper. Every AI-generated reformulation was checked by one of the authors, who supervised the tools throughout. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is theoretical and its experiments use only synthetic data; it involves no human subjects, collects no personal data, and releases no dataset or trained model.

Metrics are used as performance indicators, which makes them central in a growing number of areas, from decision-making to workplace performance evaluation (Porter, 2020). Our research can help improve the many metrics used today and better align agents whose objectives initially diverge. We are not aware of any conflict of interest.

REPRODUCIBILITY STATEMENT

To improve reproducibility, all proofs are given in the Appendix. Our assumptions are discussed in the main text where they are introduced (Section 4.1). To further support our claims, we provide in the supplementary material the ZIP file of our SymPy code (not a generative AI solution) verifying the lengthy computations of Proposition 3(ii). The code for the experiments is also provided, and Section 5.1 describes the experimental setting in full detail.

REFERENCES

Mir M Ali, NN Mikhail, and M Safiul Haq. A class of bivariate distributions including the bivariate logistic. *Journal of multivariate analysis*, 8(3):405–412, 1978.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety. *arXiv preprint arXiv:1606.06565*, 2016. doi: 10.48550/ARXIV.1606.06565.

Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, Hoda Heidari, Anson Ho, Sayash Kapoor, Leila Khalatbari, Shayne Longpre, Sam Manning, Vasilios Mavroudis, Mantas Mazeika, Julian Michael, Jessica Newman, Kwan Yee Ng, Chinasa T. Okolo, Deborah Raji, Girish Sastry, Elizabeth Seger, Theodora Skeadas, Tobin South, Emma Strubell, Florian Tramèr, Lucia Velasco, Nicole Wheeler, Daron Acemoglu, Olubayo Adekanmbi, David Dalrymple, Thomas G. Dietterich, Edward W. Felten, Pascale Fung, Pierre-Olivier Gourinchas, Fredrik Heintz, Geoffrey Hinton, Nick Jennings, Andreas Krause, Susan Leavy, Percy Liang, Teresa Ludermir, Vidushi Marda, Helen Margetts, John McDermid, Jane Munga, Arvind Narayanan, Alondra Nelson, Clara Neppel, Alice Oh, Gopal Ramchurn, Stuart Russell, Marietje Schaake, Bernhard Schölkopf, Dawn Song, Alvaro Soto, Lee Tiedrich, Gaël Varoquaux, Andrew Yao, Ya-Qin Zhang, Fahad Albalawi, Marwan Alserkal, Olubunmi Ajala, Guillaume Avrin, Christian Busch, André Carlos Ponce de Leon Ferreira de Carvalho, Bronwyn Fox, Amandeep Singh Gill, Ahmet Halit Hatip, Juha Heikkilä, Gill Jolly, Ziv Katzir, Hiroaki Kitano, Antonio Krüger, Chris Johnson, Saif M. Khan, Kyoung Mu Lee, Dominic Vincent Ligot, Oleksii Molchanovskyi, Andrea Monti, Nusu Mwamanzi, Mona Nemer, Nuria Oliver, José Ramón López Portillo, Balaraman Ravindran, Raquel Pezoa Rivera, Hammam Riza, Crystal Rugege, Ciarán Seoighe, Jerry Sheehan, Haroon Sheikh, Denise Wong, and Yi Zeng. International AI Safety Report, 2025.

Teemu Birkstedt, Matti Minkkinen, Anushree Tandon, and Matti Mäntymäki. AI governance: Themes, knowledge gaps and future agendas. *Internet Research*, 33(7):133–167, December 2023. ISSN 1066-2243. doi: 10.1108/INTR-01-2022-0042.

Alexis Boulin. A python package for sampling from copulae: clayton, 2023. URL https://arxiv.org/abs/2203.17177.

David G Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151, 1978.

Ravit Dotan, Borhane Blili-Hamelin, Ravi Madhavan, Jeanna Matthews, and Joshua Scarpino. Evolving AI Risk Management: A Maturity Model based on the NIST AI Risk Management Framework. *arXiv preprint arXiv:2401.15229*, 2024. doi: 10.48550/ARXIV.2401.15229.

El-Mahdi El-Mhamdi and Lê-Nguyên Hoang. On Goodhart's Law, with an Application to Value Alignment. *arXiv preprint arXiv:2410.09638*, 2024. doi: 10.48550/ARXIV.2410.09638.

Tom Everitt, Victoria Krakovna, Laurent Orseau, and Shane Legg. Reinforcement Learning with a Corrupted Reward Channel. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 4705–4713. International Joint Conferences on Artificial Intelligence Organization, 2017. ISBN 978-0-9992411-0-3. doi: 10.24963/ijcai.2017/656. URL https://www.ijcai.org/proceedings/2017/656.

Leo Gao, John Schulman, and Jacob Hilton. Scaling Laws for Reward Model Overoptimization. *arXiv preprint arXiv:2210.10760*, 2022. doi: 10.48550/ARXIV.2210.10760.

Charles Goodhart. Monetary relationships : A view from threadneedle street. *Papers in monetary economics 1975*, 1975.

Moritz Hardt, Nimrod Megiddo, Christos H. Papadimitriou, and Mary Wootters. Strategic classification. *CoRR,abs/1506.06980*, abs/1506.06980, 2015. URL http://arxiv.org/abs/1506.06980.

Christopher Hennessy and Charles A.E. Goodhart. Goodhart's Law and Machine Learning. *SSRN Electronic Journal*, 2020. ISSN 1556-5068. doi: 10.2139/ssrn.3639508.

Keith Hoskin. The "Awful Idea of Accountability": Inscribing People into the Measurement of Objects. In *Accountability: Power, Ethos and the Technologies of Managing*. International Thomson Business Press, London, r. munro and j. mouritsen edition, 1996.

Jennifer Hsia, Danish Pruthi, Aarti Singh, and Zachary C. Lipton. Goodhart's law applies to nlp's explanation benchmarks. *arXiv preprint arXiv:2308.14272*, 2023.

Jacek Karwowski, Oliver Hayman, Xingjian Bai, Klaus Kiendlhofer, Charlie Griffin, and Joar Max Viktor Skalse. Goodhart's law in reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5o9G4XF1LI.

Harry Kesten. Random difference equations and Renewal theory for products of random matrices. *Acta Mathematica*, 131(0):207–248, 1973. ISSN 0001-5962. doi: 10.1007/BF02392040. URL http://projecteuclid.org/euclid.acta/1485889791.

Thomas Kwa, Drake Thomas, and Adrià Garriga-Alonso. Catastrophic goodhart: regularizing RLHF with KL divergence does not mitigate heavy-tailed reward misspecification. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=UXuBzWoZGK.

Cassidy Laidlaw, Shivam Singhal, and Anca Dragan. Correlated proxies: A new definition and improved mitigation for reward hacking. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=msEr27EejF.

Robert E. Lucas. Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, 1:19–46, January 1976. ISSN 01672231. doi: 10.1016/S0167-2231(76)80003-6.

David Manheim. Building less-flawed metrics: Understanding and creating better measurement and incentive systems. *Patterns*, 4(10):100842, October 2023. ISSN 26663899. doi: 10.1016/j.patter.2023.100842.

David Manheim and Scott Garrabrant. Categorizing Variants of Goodhart's Law. *arXiv preprint arXiv:1803.04585*, 2018. doi: 10.48550/ARXIV.1803.04585.

Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, AmiT Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E. Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. SymPy: Symbolic computing in Python. *PeerJ Computer Science*, 3:e103, January 2017. ISSN 2376-5992. doi: 10.7717/peerj-cs.103.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=JYtwGwIL7ye.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7599–7609. PMLR, 2020. URL https://proceedings.mlr.press/v119/perdomo20a.html.

Theodore M. Porter. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press, 2020. ISBN 978-0-691-20841-1. doi: 10.23943/princeton/9780691208411.001.0001. URL https://doi.org/10.23943/princeton/9780691208411.001.0001.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and Characterizing Reward Hacking. *arXiv preprint arXiv:2209.13085*, 2022. doi: 10.48550/ARXIV.2209.13085.

James E. Smith and Robert L. Winkler. The Optimizer's Curse: Skepticism and Postdecision Surprise in Decision Analysis. *Management Science*, 52(3):311–322, 2006. ISSN 0025-1909, 1526-5501. doi: 10.1287/mnsc.1050.0451. URL https://pubsonline.informs.org/doi/10.1287/mnsc.1050.0451.

Marilyn Strathern. 'Improving ratings': Audit in the British University system. *European Review*, 5(3):305–321, July 1997. ISSN 10627987, 1234981X. doi: 10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4.

Rachel L. Thomas and David Uminsky. Reliance on metrics is a fundamental challenge for AI. *Patterns*, 3(5):100476, May 2022. ISSN 26663899. doi: 10.1016/j.patter.2022.100476.

Simon Zhuang and Dylan Hadfield-Menell. Consequences of misaligned AI. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15763–15773. Curran Associates, Inc., 2020.

Supplementary material

A Explicit couplings

A.1 Gaussian goal and Gaussian discrepancy

Throughout this section, $(G, \xi) \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = \begin{pmatrix} a & c \\ c & b \end{pmatrix}$, $a, b > 0$ and $ab > c^2$, and $M = G + \xi$. We write $s^2 := \mathrm{Var}(M) = a + b + 2c$, which is positive since Σ is positive definite, and φ and $\bar{\Phi}$ for the density and survival function of $\mathcal{N}(0, 1)$.

Reduction to a regression on M. Let $\beta := \mathrm{Cov}(G, M)/\mathrm{Var}(M) = (a+c)/s^2$ and $\varepsilon := G - \beta M$. The pair (ε, M) is Gaussian with $\mathrm{Cov}(\varepsilon, M) = 0$, so ε is independent of M, centered, with variance

$$v := \mathrm{Var}(\varepsilon) = a - \frac{(a+c)^2}{s^2} = \frac{ab - c^2}{s^2} > 0.$$

Conditionally on $\{M > m\}$, ε is still independent of M and keeps its distribution. Writing $\sigma_m^2 := \mathrm{Var}(M \mid M > m)$, we therefore get

$$\mathbb{E}[G \mid M > m] = \beta\, \mathbb{E}[M \mid M > m], \tag{1}$$
$$\mathrm{Cov}(G, M \mid M > m) = \beta\, \sigma_m^2, \qquad \mathrm{Var}(G \mid M > m) = \beta^2 \sigma_m^2 + v,$$

hence

$$\mathrm{Corr}(G, M \mid M > m) = \frac{\beta\, \sigma_m}{\sqrt{\beta^2 \sigma_m^2 + v}}. \tag{2}$$

Likewise, $\xi = (1 - \beta)M - \varepsilon$ with $1 - \beta = (b + c)/s^2$, so that $\mathbb{E}[\xi \mid M > m] = (1 - \beta)\, \mathbb{E}[M \mid M > m]$. Everything thus reduces to the first two moments of a truncated Gaussian.

Lemma 4 (Truncated standard Gaussian)**.** *Let $Z \sim \mathcal{N}(0, 1)$. As $t \to \infty$,*

$$\bar{\Phi}(t) = \frac{\varphi(t)}{t}\Big(1 - \frac{1}{t^2} + \frac{3}{t^4} + O(t^{-6})\Big),$$
$$\mathbb{E}[Z \mid Z > t] = t + \frac{1}{t} - \frac{2}{t^3} + \frac{10}{t^5} + O(t^{-7}),$$
$$\mathrm{Var}(Z \mid Z > t) = \frac{1}{t^2} - \frac{6}{t^4} + O(t^{-6}).$$

Proof. Since $\varphi'(u) = -u\varphi(u)$, repeated integrations by parts of $\bar{\Phi}(t) = \int_t^\infty \varphi(u)\, du$ give the classical expansion of Mills' ratio,

$$\frac{\bar{\Phi}(t)}{\varphi(t)} = \frac{1}{t} - \frac{1}{t^3} + \frac{3}{t^5} - \frac{15}{t^7} + O(t^{-9}),$$

which is the first claim. Next, $\lambda(t) := \mathbb{E}[Z \mid Z > t] = \varphi(t)/\bar{\Phi}(t)$, and inverting the previous expansion yields the second claim. Finally, one more integration by parts gives $\int_t^\infty u^2 \varphi(u)\, du = t\varphi(t) + \bar{\Phi}(t)$, that is, $\mathbb{E}[Z^2 \mid Z > t] = 1 + t\lambda(t)$, so that $\mathrm{Var}(Z \mid Z > t) = 1 - \lambda(t)\big(\lambda(t) - t\big)$. Substituting the second claim gives the third. □

Proof of Proposition 3(i). Write $M = sZ$ with $Z \sim \mathcal{N}(0, 1)$ and apply Lemma 4 with $t = m/s$:

$$\mathbb{P}(M > m) = \frac{s}{\sqrt{2\pi}\, m}\, e^{-m^2/(2s^2)}\Big(1 - \frac{s^2}{m^2} + \frac{3s^4}{m^4} + O(m^{-6})\Big),$$
$$\mathbb{E}[M \mid M > m] = m + \frac{s^2}{m} - \frac{2s^4}{m^3} + O(m^{-5}),$$
$$\sigma_m^2 = \frac{s^4}{m^2} - \frac{6s^6}{m^4} + O(m^{-6}).$$

By (1),

$$\mathbb{E}[G \mid M > m] = \frac{a + c}{a + b + 2c}\Big(m + \frac{s^2}{m} - \frac{2s^4}{m^3} + O(m^{-5})\Big) \underset{m \to \infty}{\sim} \frac{a + c}{a + b + 2c}\, m.$$

For the correlation, $\sigma_m = \frac{s^2}{m}\big(1 - \frac{3s^2}{m^2} + O(m^{-4})\big) \to 0$ while $v > 0$ is fixed, so expanding (2) gives

$$\mathrm{Corr}(G, M \mid M > m) = \frac{(a+c)\sqrt{a+b+2c}}{\sqrt{ab-c^2}}\,\frac{1}{m}\Big(1 - \Big(3 + \frac{(a+c)^2}{2(ab-c^2)}\Big)\frac{a+b+2c}{m^2} + O(m^{-4})\Big).$$

In particular, if $a + c > 0$, then $\mathbb{E}[G \mid M > m] \to +\infty = \sup S_G$ while $\mathrm{Corr}(G, M \mid M > m) \to 0$: this is benign Goodhart. □

Other signs of $a + c$. The computation above never uses the sign of $a + c = \mathrm{Cov}(G, M)$. If $a + c < 0$, then $\beta < 0$ and $\mathbb{E}[G \mid M > m] \sim \beta m \to -\infty = \inf S_G$: strong Goodhart. If $a + c = 0$, then $G = \varepsilon$ is independent of M and $\mathbb{E}[G \mid M > m] = 0$ for every m: weak Goodhart.

A.2 Exponential goal and heavy-tailed discrepancy

In this case, the goal follows an exponential distribution. The conditional distribution of the discrepancy given the goal is (for $b \in (1, \infty)$, $\eta \in (0, \infty)$)

$$p_{\xi|G}(u) = G\exp(-G((\frac{u}{\eta})^{b-1} - 1))\frac{u^{b-2}}{\eta^{b-1}}(b-1)\mathbb{1}\{u > \eta\}.$$

The discrepancy thus defined follows a power law with shape parameter b and scale parameter η.

A.3 Lemmas

We need the following lemmas, which are used in nearly all subsequent proofs.

Lemma 5. *Let $P, Q \in C^\infty(I)$ with $Q \neq 0$ on the interval I, and set $f_Q(P) := (P/Q)'$. For every $n \in \mathbb{N}$,*

$$f_Q^n(P) = \sum_{k=0}^{n} \sum_{i_0+\cdots+i_k=n-k} c^{(n)}_{i_0,\ldots,i_k}\,\frac{P^{(i_0)}\,Q^{(i_1+1)}\cdots Q^{(i_k+1)}}{Q^{\,n+k}},$$

where the $c^{(n)}_{i_0,\ldots,i_k}$ are integers of sign $(-1)^k$ that depend only on n and on the indices, not on P, Q or x.

Proof. By induction on n, the case $n = 0$ being $f_Q^0(P) = P$. Take a term of $f_Q^n(P)$ with k factors of Q and divide it by Q: its denominator becomes Q^{n+k+1}. Differentiating it then produces two kinds of terms. If the derivative hits $P^{(i_0)}$ or one of the $Q^{(i_j+1)}$, one index increases by one: the indices now sum to $(n+1) - k$ over the denominator $Q^{(n+1)+k}$. If it hits the denominator, it produces $-(n+k+1)\,Q'/Q^{n+k+2}$, that is, a new factor $Q^{(0+1)}$: there are now $k+1$ factors, whose indices sum to $(n+1)-(k+1)$, over $Q^{(n+1)+(k+1)}$. In both cases the term has the announced form at rank $n+1$, and its coefficient is an integer that depends only on the indices, whose sign changes exactly when a factor of Q is created. □

Before the main lemma, we compute the derivatives of Q of all orders.

Lemma 6. *Let $b > 1$, $\eta > 0$, $m > 0$, and $Q(g) := -g\big(\frac{m-g}{\eta}\big)^{b-1}$ for $g < m$. For every $n \geq 1$,*

$$Q^{(n)}(g) = (-1)^n\,\frac{(nm - bg)\,(m-g)^{b-n-1}}{\eta^{b-1}}\,\prod_{i=1}^{n-1}(b-i).$$

In particular, when b is an integer, $Q^{(b)} \equiv (-1)^b\,b!/\eta^{b-1}$ and $Q^{(n)} \equiv 0$ for every $n > b$.

Proof. Since $\frac{d^k}{dg^k}(m-g)^{b-1} = (-1)^k\prod_{i=1}^{k}(b-i)\,(m-g)^{b-1-k}$ and all derivatives of g of order at least two vanish, the Leibniz rule gives

$$-\eta^{b-1}Q^{(n)}(g) = (-1)^n\prod_{i=1}^{n}(b-i)\,g\,(m-g)^{b-1-n} + n\,(-1)^{n-1}\prod_{i=1}^{n-1}(b-i)\,(m-g)^{b-n}.$$

Factoring out $(-1)^{n-1}\prod_{i=1}^{n-1}(b-i)\,(m-g)^{b-n-1}$ leaves $n(m-g) - (b-n)g = nm - bg$, which gives the claim. □

Lemma 7. *Let $l \in \mathbb{N}$, $s \geq 0$, $b > 1$ and $\eta > 0$. Define $Q(g) := -g\left(\frac{m-g}{\eta}\right)^{b-1}$ and $P_l(g) := \frac{g^l}{(m-g)^s}$. Then, for every $N \in \mathbb{N}$,*

$$\int_0^{m-\eta} \frac{g^l}{(m-g)^s} \exp(-g\left(\frac{m-g}{\eta}\right)^{b-1})dg \underset{m\to\infty}{=} \sum_{n=0}^{N}(-1)^n \left[\frac{f_{Q'}^n(P_l(g))}{Q'(g)} \exp(Q(g))\right]_0^{\frac{m}{b+1}} + o(\frac{1}{m^{(N+1)b-l+s-1}}).$$

Proof. We first split the integral in two:

$$\int_0^{m-\eta} \frac{g^l}{(m-g)^s} \exp(-g\left(\frac{m-g}{\eta}\right)^{b-1})dg$$
$$= \underbrace{\int_0^{m/(b+1)} \frac{g^l}{(m-g)^s} \exp(-g\left(\frac{m-g}{\eta}\right)^{b-1})dg}_{I_1} + \underbrace{\int_{m/(b+1)}^{m-\eta} \frac{g^l}{(m-g)^s} \exp(-g\left(\frac{m-g}{\eta}\right)^{b-1})dg}_{I_2}.$$

We bound I_2 crudely, since it is negligible:

$$\int_{m/(b+1)}^{m-\eta} \frac{g^l}{(m-g)^s} \exp(-g\left(\frac{m-g}{\eta}\right)^{b-1})dg \leq \frac{(m-\eta)^{l+1}}{\eta^s} \exp(-\frac{m}{b+1}).$$

Write $J := [0, m/(b+1)]$. On J we have $m - g \in [\frac{b}{b+1}m,\, m]$ and $m - bg \in [\frac{1}{b+1}m,\, m]$, so that $Q'(g) = -(m-bg)(m-g)^{b-2}/\eta^{b-1}$ does not vanish. Since $\int_J P\, e^Q\, dg = \left[\frac{P}{Q'}\, e^Q\right]_0^{m/(b+1)} - \int_J f_{Q'}(P)\, e^Q\, dg$, iterating $N+1$ times (with $f_{Q'}^0$ the identity) gives

$$\int_J P_l\, e^Q\, dg = \sum_{n=0}^{N}(-1)^n \Big[\frac{f_{Q'}^n(P_l)}{Q'}\, e^Q\Big]_0^{m/(b+1)} + (-1)^{N+1} \int_J f_{Q'}^{N+1}(P_l)\, e^Q\, dg.$$

It remains to bound the last integral. In what follows, C denotes a constant depending only on b, η, l, s, N (neither on m nor on g), which may change from line to line. For $g \in J$:

$$|Q'(g)| \geq C^{-1} m^{b-1} \tag{3}$$
$$|Q^{(i+2)}(g)| \leq C\, m\, (m-g)^{b-i-3} \leq C\, m^{b-i-2}, \forall\, i \geq 0 \tag{4}$$
$$|P_l^{(j)}(g)| \leq C\, m^{l-s-j} \tag{5}$$
$$Q(g) = -\, g\big(\frac{m-g}{\eta}\big)^{b-1} \leq -g\,\big(\frac{bm}{(b+1)\eta}\big)^{b-1}. \tag{6}$$

Equation (3) holds since $m - bg \geq \frac{m}{b+1}$ and $(m-g)^{b-2} \geq C^{-1} m^{b-2}$. Equation (4) is a consequence of Lemma 6. Equation (5) follows from the Leibniz rule applied to P_l:

$$P_l^{(j)}(g) = \sum_{q=0}^{\min(j,l)} \binom{j}{q} \frac{l!}{(l-q)!}\, g^{l-q}\, s(s+1)\cdots(s+j-q-1)\, (m-g)^{-s-j+q}.$$

Equation (6) follows from $m - g \geq \frac{b}{b+1}m$ and $b > 1$.

By Lemma 5 applied with Q' in place of Q, $f_{Q'}^{N+1}(P_l)$ is a finite sum, with coefficients depending only on N, of terms $P_l^{(i_0)}\, Q^{(i_1+2)} \cdots Q^{(i_k+2)}/(Q')^{N+1+k}$ with $i_0 + \cdots + i_k = N + 1 - k$. By (3)–(5), each of them is bounded on J by

$$C\, m^{l-s-i_0} \prod_{j=1}^{k} m^{b-i_j-2}\; m^{-(b-1)(N+1+k)} = C\, m^{\,l-s-b(N+1)},$$

where the exponents simplify thanks to $i_0 + \cdots + i_k = N + 1 - k$. Together with (6),

$$\Big|\int_J f_{Q'}^{N+1}(P_l)\, e^Q\, dg\Big| \leq C\, m^{\,l-s-b(N+1)} \int_0^{\infty} e^{-g\,(bm/((b+1)\eta))^{b-1}}\, dg = O\big(m^{-((N+2)b-l+s-1)}\big),$$

which is $o\big(m^{-((N+1)b-l+s-1)}\big)$. Finally, by (6), the boundary terms at $g = m/(b+1)$ in the sum are $O\big(m^C e^{-m^b/C}\big)$, hence negligible. This concludes the proof, for every real $b > 1$.

□

The following lemmas, all consequences of Lemma 7, are the building blocks for the computations below. Note that all integrals below are well defined and finite for $b \in (1, \infty)$ and $\eta \in (0, \infty)$.

Lemma 8.

$$\int_0^{m-\eta} \exp(-g\left(\frac{m-g}{\eta}\right)^{b-1})dg \underset{m\to\infty}{=} \frac{\eta^{b-1}}{m^{b-1}} + 2\frac{\eta^{2b-2}\,(b-1)}{m^{2b-1}} + \frac{3\eta^{3b-3}\,(b-1)\,(3b-2)}{m^{3b-1}} + 8\frac{\eta^{4b-4}\,(b-1)\,(2b-1)\,(4b-3)}{m^{4b-1}} + o\left(\frac{1}{m^{4b-1}}\right).$$

Proof. Apply Lemma 7 with $l = s = 0$ and $N = 3$. □

Lemma 9.

$$\int_0^{m-\eta} g\exp(-g\left(\frac{m-g}{\eta}\right)^{b-1})dg \underset{m\to\infty}{=} \frac{\eta^{2b-2}}{m^{2b-2}} + \frac{6\,(b-1)\,\eta^{3b-3}}{m^{3b-2}} + \frac{12\eta^{4b-4}\,(b-1)\,(4b-3)}{m^{4b-2}} + o\left(\frac{1}{m^{4b-2}}\right).$$

Proof. Apply Lemma 7 with $l = 1$, $s = 0$ and $N = 3$. □

Lemma 10.

$$\int_0^{m-\eta} g^2\exp(-g\left(\frac{m-g}{\eta}\right)^{b-1})dg \underset{m\to\infty}{=} \frac{2\eta^{3b-3}}{m^{3b-3}} + \frac{24\eta^{4b-4}\,(b-1)}{m^{4b-3}} + o\left(\frac{1}{m^{4b-3}}\right).$$

Proof. Apply Lemma 7 with $l = 2$, $s = 0$ and $N = 3$. □

Lemma 11.

$$\int_0^{m-\eta} \frac{\eta^{b-1}}{(m-u)^{b-1}}\exp(-u\left(\frac{m-u}{\eta}\right)^{b-1})du \underset{m\to\infty}{=} \frac{\eta^{2b-2}}{m^{2b-2}} + \frac{3\eta^{3b-3}\,(b-1)}{m^{3b-2}} + \frac{4\eta^{4b-4}\,(b-1)\,(4b-3)}{m^{4b-2}} + o\left(\frac{1}{m^{4b-2}}\right).$$

Proof. Apply Lemma 7 with $l = 0$, $s = b-1$ and $N = 3$. □

Lemma 12.

$$\int_0^{m-\eta} \left(\frac{\eta^{b-1}}{(m-u)^{b-2}}\right) exp(-u\left(\frac{m-u}{\eta}\right)^{b-1})du \underset{m\to\infty}{=} \frac{\eta^{2b-2}}{m^{2b-3}} - \frac{\eta^{3b-3}\,(4-3b)}{m^{3b-3}} + \frac{4\eta^{4b-4}\,(b-1)\,(4b-5)}{m^{4b-3}} + \frac{5\eta^{5b-5}\,(b-1)\,(5b-6)\,(5b-4)}{m^{5b-3}} + o\left(\frac{1}{m^{5b-3}}\right).$$

Proof. Apply Lemma 7 with $l = 0$, $s = b-2$ and $N = 3$. □

Lemma 13.

$$\int_0^{m-\eta} \left(\frac{\eta^{2b-2}}{(m-u)^{2b-2}}\right) exp(-u\left(\frac{m-u}{\eta}\right)^{b-1})du \underset{m\to\infty}{=} \frac{\eta^{3b-3}}{m^{3b-3}} + \frac{4\eta^{4b-4}\,(b-1)}{m^{4b-3}} + \frac{5\eta^{5b-5}\,(b-1)\,(5b-4)}{m^{5b-3}} + \frac{12\,(b-1)\,(3b-2)\,(6b-5)\,\eta^{6b-6}}{m^{6b-3}} + o\left(\frac{1}{m^{6b-3}}\right).$$

Proof. Apply Lemma 7 with $l = 0$, $s = 2b-2$ and $N = 3$. □

Lemma 14. *If $G \sim \mathcal{E}(1)$ and the conditional density of ξ is $p_{\xi|G}(x) := G\exp(-G((\frac{x}{\eta})^{b-1} - 1))\frac{x^{b-2}}{\eta^{b-1}}(b-1)\mathbb{1}\{x > \eta\}$, then*

$$\mathbb{P}\left(G+\xi > m\right) \underset{m\to\infty}{=} \frac{\eta^{b-1}}{m^{b-1}} + \frac{2\eta^{2b-2}(b-1)}{m^{2b-1}} + \frac{3\eta^{3b-3}(b-1)(3b-2)}{m^{3b-1}} + \frac{8\eta^{4b-4}(b-1)(2b-1)(4b-3)}{m^{4b-1}} + o\left(\frac{1}{m^{4b-1}}\right).$$

Proof.

$$\begin{aligned}
\mathbb{P}\left(G+\xi > m\right) &= \int_{\mathbb{R}} \exp(-g) \int_{\mathbb{R}} \exp(-g(\left(\frac{x}{\eta}\right)^{b-1} - 1)x^{b-2}\frac{g}{\eta^{b-1}}(b-1)\mathbb{1}\{x > \eta, g > 0, g + x > m\}dxdg \\
&= \int_0^{\infty} \exp(-g) \int_{\mathbb{R}} \exp(-g(\left(\frac{x}{\eta}\right)^{b-1} - 1)x^{b-2}\frac{g}{\eta^{b-1}}(b-1)\mathbb{1}\{x > \eta, x > m - g\}dxdg \\
&= \int_0^{\infty} \exp(-g) \int_{\mathbb{R}} \exp(-g(\left(\frac{x}{\eta}\right)^{b-1} - 1)x^{b-2}\frac{g}{\eta^{b-1}}(b-1)\mathbb{1}\{x > \max(\eta, m-g)\}dxdg,
\end{aligned}$$

as $g > m - \eta \implies \eta > m - g$

$$= \int_0^{m-\eta} \exp(-g) \underbrace{\int_{m-g}^{\infty} \exp(-g(\left(\frac{x}{\eta}\right)^{b-1} - 1)x^{b-2}\frac{g}{\eta^{b-1}}(b-1)dx}_{A_1} dg + S_G(m-\eta).$$

For the survival function of G:

$$\begin{aligned}
S_G(m-\eta) &= \int_{m-\eta}^{\infty} \exp(-g)\, dg \\
&= \exp(-m)\exp(\eta) \\
&= o(\frac{1}{m^{2b}}).
\end{aligned}$$

For A_1:

$$\begin{aligned}
A_1 = \int_{m-g}^{\infty} \exp(-g\left(\frac{x}{\eta}\right)^{b-1} x^{b-2}\frac{g}{\eta^{b-1}}(b-1)dx &\underset{b>1}{=} \left[\exp(-g\left(\frac{x}{\eta}\right)^{b-1})\right]_{m-g}^{\infty} \\
&\underset{b>1}{=} \exp(-g\left(\frac{m-g}{\eta}\right)^{b-1}).
\end{aligned}$$

Hence

$$\mathbb{P}\left(G+\xi > m\right) \underset{m\to\infty}{=} \int_0^{m-\eta} \exp(-g\left(\frac{m-g}{\eta}\right)^{b-1})dg + o(\frac{1}{m^{2b}}).$$

Applying Lemma 8 then yields the result.

□

A.4 Conditional expectation of the goal in Proposition 3(ii)

Lemma 15.

$$\mathbb{E}\left[G|M > m\right] \underset{m\to\infty}{=} \frac{\eta^{b-1}}{m^{b-1}} + \frac{4\eta^{2b-2}(b-1)}{m^{2b-1}} + \frac{\eta^{3b-3}(b-1)(31b-22)}{m^{3b-1}} + o\left(\frac{1}{m^{3b-1}}\right).$$

Proof. Let $\zeta := \mathbb{P}(G + \xi > m)$. Then

$$\begin{aligned}\zeta\mathbb{E}[G|G+\xi>m] &= \int_0^{m-\eta} \exp(-g) g \int_{m-g}^{\infty} \exp(-g(\left(\frac{x}{\eta}\right)^{b-1} - 1)) x^{b-2} \frac{g}{\eta^{b-1}}(b-1)dxdg \\ &\quad + \int_{m-\eta}^{\infty} \exp(-g) g dg \\ &= \underbrace{\int_0^{m-\eta} g\exp(-g\left(\frac{m-g}{\eta}\right)^{b-1})dg}_{A_2} + \underbrace{\int_{m-\eta}^{\infty} g\exp(-g)dg}_{A_3}.\end{aligned}$$

Focusing first on A_3:

$$\begin{aligned}\int_{m-\eta}^{\infty} g\exp(-g)dg &= [-g\exp(-g)]_{m-\eta}^{\infty} + \int_{m-\eta}^{\infty}\exp(-g)dg \\ &= (m-\eta)\exp(-(m-\eta)) + \exp(-(m-\eta)) \\ &= \exp(-m)\exp(\eta)(m-\eta+1).\end{aligned}$$

Lemma 9 applies to A_2.

It remains to expand the normalizing factor:

$$\begin{aligned}&\frac{1}{\zeta} \\ &\underset{m\to\infty}{=} \frac{1}{\frac{\eta^{b-1}}{m^{b-1}} + \frac{2\eta^{2b-2}(b-1)}{m^{2b-1}} + \frac{3\eta^{3b-3}(b-1)(3b-2)}{m^{3b-1}} + \frac{8\eta^{4b-4}(b-1)(2b-1)(4b-3)}{m^{4b-1}} + o\left(\frac{1}{m^{4b-1}}\right)} \\ &\underset{m\to\infty}{=} \frac{m^{b-1}}{\eta^{b-1}(1 + 2\eta^{b-1}m^{-b}(b-1) + 3\eta^{2b-2}m^{-2b}(b-1)(3b-2) + 8\eta^{3b-3}m^{-3b}(b-1)(2b-1)(4b-3) + o\left(\frac{1}{m^{3b}}\right))},\end{aligned}$$

using the expansion of $1/(1+x)$:

$$\begin{aligned}&\underset{m\to\infty}{=} \frac{m^{b-1}}{\eta^{b-1}}(1 - \frac{2\eta^{b-1}(b-1)}{m^b} - \frac{\eta^{2b-2}(5b-2)(b-1)}{m^{2b}} - \frac{4\eta^{3b-3}(b-1)(3b-2)(3b-1)}{m^{3b}} + O\left(\frac{1}{m^{4b}}\right)) \\ &\underset{m\to\infty}{=} \frac{m^{b-1}}{\eta^{b-1}} - \frac{2(b-1)}{m} - \frac{\eta^{b-1}(b-1)(5b-2)}{m^{b+1}} - \frac{4\eta^{2b-2}(b-1)(3b-2)(3b-1)}{m^{2b+1}} + O\left(\frac{1}{m^{3b+1}}\right) \\ &=: \Delta\end{aligned}$$

We denote this expansion by Δ, as it is used several times below. Multiplying the expansion of Lemma 9 by Δ yields the result. □

A.5 Conditional expectation of ξ in Proposition 3(ii)

In the following, we denote $\zeta := \mathbb{P}(G + \xi > m)$.

Lemma 16.

$$\begin{aligned}\mathbb{E}[\xi|M>m] &\underset{m\to\infty}{=} m\frac{b-1}{b-2} - \frac{2\eta^{b-1}(b-1)}{m^{b-1}(b-2)} - \frac{8\eta^{2b-2}(b-1)^2}{m^{2b-1}(b-2)} - \frac{2\eta^{3b-3}(b-1)^2(31b-22)}{m^{3b-1}(b-2)} \\ &\quad + o\left(\frac{1}{m^{3b-1}}\right).\end{aligned}$$

Proof.

$$\zeta\mathbb{E}[\xi|G+\xi>m] = \int_0^{\infty}\int_{\mathbb{R}} \exp(-g(\left(\frac{x}{\eta}\right)^{b-1})x^{b-1}\frac{g}{\eta^{b-1}}(b-1)\mathbb{1}\{x > \max(\eta, m-g)\}dxdg$$

denoting $\mathcal{M}(x) := max(0, m-x)$

$$= \int_\eta^\infty \frac{x^{b-1}(b-1)}{\eta^{b-1}} \underbrace{\int_{\mathcal{M}(x)}^{+\infty} g \exp(-g(\left(\frac{x}{\eta}\right)^{b-1})dg}_{A_4} dx.$$

We first compute A_4 by integration by parts:

$$\int_{\mathcal{M}(x)}^{+\infty} g \exp(-g(\left(\frac{x}{\eta}\right)^{b-1})dg = \mathcal{M}(x) \left(\frac{\eta}{x}\right)^{b-1} \exp(-\mathcal{M}(x) \left(\frac{x}{\eta}\right)^{b-1}) + \int_{\mathcal{M}(x)}^{+\infty} \frac{\eta^{b-1}}{x^{b-1}} \exp(-g \left(\frac{x}{\eta}\right)^{b-1})dg.$$

Plugging this into the full integral:

$$\begin{aligned}
&\zeta \mathbb{E}\left[\xi | G + \xi > m\right] \\
&= \int_\eta^\infty (b-1)\mathcal{M}(x) \exp(-\mathcal{M}(x) \left(\frac{x}{\eta}\right)^{b-1}) + (b-1) \int_{\mathcal{M}(x)}^{+\infty} \exp(-g \left(\frac{x}{\eta}\right)^{b-1})dgdx \\
&= \int_\eta^\infty (b-1)\mathcal{M}(x) \exp(-\mathcal{M}(x) \left(\frac{x}{\eta}\right)^{b-1}) + (b-1)\frac{\eta^{b-1}}{x^{b-1}} \exp(-\mathcal{M}(x) \left(\frac{x}{\eta}\right)^{b-1})dx \\
&= \int_\eta^m (b-1)(m-x) \exp(-(m-x) \left(\frac{x}{\eta}\right)^{b-1}) + (b-1)\frac{\eta^{b-1}}{x^{b-1}} \exp(-(m-x) \left(\frac{x}{\eta}\right)^{b-1})dx \\
&\quad + \int_m^\infty (b-1)\frac{\eta^{b-1}}{x^{b-1}} dx \\
&= \underbrace{\int_0^{m-\eta} (b-1)u \exp(-u \left(\frac{m-u}{\eta}\right)^{b-1})du}_{A_5} + \underbrace{\int_0^{m-\eta} (b-1)\frac{\eta^{b-1}}{(m-u)^{b-1}} \exp(-u \left(\frac{m-u}{\eta}\right)^{b-1})du}_{A_6} \\
&\quad + \frac{(b-1)\eta^{b-1}}{(b-2)m^{b-2}}.
\end{aligned}$$

Lemmas 9 and 11 give asymptotic expansions of A_5 and A_6, respectively. Summing yields the unnormalized expectation $\zeta \mathbb{E}\left[\xi | M > m\right]$, which, multiplied by Δ, yields the result. □

B General results

B.1 General tools

Lemma 17. *Let l be a slowly varying function, positive on a neighborhood of $+\infty$. Then for any $u : x \mapsto u(x)$ with $u(x) \underset{x\to+\infty}{\to} +\infty$ and any $z > 0$,*

$$l(u(x)) \underset{x\to+\infty}{=} o\left(u(x)^z\right).$$

Proof. Fix $z > 0$. By Karamata's representation theorem there exist $B > 0$, a bounded function ψ and a function δ with $\delta(t) \underset{t\to\infty}{\to} 0$ such that for $t \geq B$,

$$l(t) = \psi(t) \exp\left(\int_B^t \frac{\delta(s)}{s} ds\right).$$

Since $\delta(s) \to 0$, there exists $C_z \geq B$ such that $\delta(s) < z/2$ for all $s > C_z$. Splitting the integral at C_z,

$$l(t) = \psi(t) \exp\left(\int_B^{C_z} \frac{\delta(s)}{s} ds\right) \exp\left(\int_{C_z}^t \frac{\delta(s)}{s} ds\right)$$

using that ψ is bounded and that the middle factor is a constant depending only on z, which we absorb into C'_z:

$$\leq C'_z \exp\left(\frac{z}{2}\int_{C_z}^{t}\frac{ds}{s}\right) \;=\; C'_z\left(\frac{t}{C_z}\right)^{z/2} \;=\; O\left(t^{z/2}\right).$$

As $t^{z/2} = o(t^z)$, composing with $u(x) \to +\infty$ gives the result. □

The next lemma is the workhorse. It bounds an expectation involving the *conditional* tail of ξ by its *unconditional* tail, at the cost of a truncation of W. No assumption is made on the tail of ξ, which is why the same statement serves both theorems.

Lemma 18 (Truncation bound)**.** *Let ξ and G be random variables and let $W \geq 0$ be $\sigma(G)$-measurable with $\mathbb{E}[W^p] < \infty$ for some $p > 1$. Then for every $z > 0$ and every $m > 0$,*

$$\mathbb{E}\left[W\mathbb{P}\left(\xi > m|G\right)\right] \;\leq\; m^z\,\mathbb{P}\left(\xi > m\right) \;+\; \frac{\mathbb{E}\left[W^p\right]}{m^{z(p-1)}}.$$

Proof. Split the expectation on $\{W \leq m^z\}$ and its complement. On the first event $W \leq m^z$, and the tower property gives $\mathbb{E}\left[\mathbb{P}\left(\xi > m|G\right)\right] = \mathbb{P}\left(\xi > m\right)$. On the second, $(W/m^z)^{p-1} \geq 1$ yields $W\mathbb{1}\{W > m^z\} \leq W^p m^{z(1-p)}$, and we bound $\mathbb{P}\left(\xi > m|G\right) \leq 1$. □

The two terms pull in opposite directions: a large z weakens the truncation but makes the residual moment term decay faster. The following lemma records the three regimes we need. Note that the conclusions of (i) and (ii) are the *same*: the light- and heavy-tailed discrepancy cases differ only in the moment required of W.

Lemma 19 (Regimes)**.** *Let $1/\gamma > 1$ and let $W \geq 0$ be $\sigma(G)$-measurable.*

(i) If ξ is right light-tailed and $\mathbb{E}[W^p] < \infty$ for some $p > 1$, then $\mathbb{E}\left[W\mathbb{P}\left(\xi > m|G\right)\right] = o\left(m^{1-1/\gamma}\right)$.

(ii) If $\mathbb{P}\left(\xi > m\right) = l_\xi(m)m^{-1/\alpha}$ with l_ξ slowly varying, $1/\alpha - 1/\gamma > 0$ and $\mathbb{E}\left[W^{1/\gamma}\right] < \infty$, then $\mathbb{E}\left[W\mathbb{P}\left(\xi > m|G\right)\right] = o\left(m^{1-1/\gamma}\right)$.

(iii) If W is bounded, then in either setting $\mathbb{E}\left[W\mathbb{P}\left(\xi > m|G\right)\right] = o\left(m^{-1/\gamma}\right)$.

Proof. Throughout, $\nu := 1/\alpha$ and $\tau := 1/\gamma$. All exponents below are reached strictly, which absorbs the slowly varying factor carried by the leading terms.

For (iii), since W is bounded, the tower property gives

$$\mathbb{E}\left[W\mathbb{P}\left(\xi > m|G\right)\right] \leq \|W\|_\infty \mathbb{P}\left(\xi > m\right).$$

This is $O(e^{-sm})$ in the light-tailed case. In the heavy-tailed case Lemma 17 gives $O(m^{q-\nu})$ for every $q > 0$, and, since $\nu - \tau > 0$, we may take $q < \nu - \tau$.

For (i) and (ii), apply Lemma 18. In case (i) the first term is $O(m^z e^{-sm})$, negligible for any z, so only $-z(p-1) < 1 - \tau$ constrains the choice; as $\tau > 1$ this holds for large z, whatever $p > 1$.

In case (ii), Lemma 17 makes the first term $o(m^{z+q-\nu})$ for every $q > 0$, so both terms constrain z. The exponents $z - \nu$ and $-z(p-1)$ being respectively increasing and decreasing in z, their maximum is minimized where they coincide, at $z^\star = \frac{\nu}{p}$, both then equaling $-\nu\frac{p-1}{p}$. Hence $\mathbb{E}\left[W\mathbb{P}\left(\xi > m|G\right)\right] = o(m^\beta)$ for every $\beta > -\nu\frac{p-1}{p}$, and such a β falls below $1 - \tau$ as soon as

$$-\nu\frac{p-1}{p} < 1 - \tau \iff \frac{\nu}{p} < \nu - \tau + 1 \iff p > p^\star := \frac{\nu}{\nu - \tau + 1},$$

using $\nu - \tau + 1 > 1 > 0$. Finally, $p^\star < \tau$ rewrites as $(\nu - \tau)(\tau - 1) > 0$, true since $\nu > \tau > 1$, so $p = 1/\gamma$ is admissible. □

B.2 THEOREMS

Both theorems below follow the same architecture: the integral defining $\mathbb{E}[G|M > m]$ is split at 0, the negative part is shown to be negligible, and the positive part is cut at a threshold $m + t_m$ whose choice balances the coverage of the support of ξ against the decay of $\mathbb{P}(G > m + t_m)$.

The only difference between the light- and heavy-tailed discrepancy cases lies in how fast t_m may be taken to grow.

Theorem 1. *Let G be heavy-tailed with tail index γ, where $1/\gamma > 1$, and slowly varying function l, with $\mathbb{E}[|G|^p \mathbb{1}\{G < 0\}] < \infty$ for some $p > 1$. Let ξ be right and left light-tailed, where $s > 0$ is an exponent for which ξ satisfies both conditions of Definition 3, and set $M := G + \xi$. Then, for any $0 < c < s$,*

$$\mathbb{E}[G|M > m] \underset{m\to\infty}{\geq} m + o(m).$$

Proof. Splitting the integral at 0:

$$\begin{aligned}&\int_{\mathbb{R}} g p_G(g) \int_{x\geq m-g} p_{\xi|G}(x)dxdg \\ =&\underbrace{\int_{\mathbb{R}^-} g p_G(g) \int_{x\geq m-g} p_{\xi|G}(x)dxdg}_{I_1} + \underbrace{\int_{\mathbb{R}^+} g p_G(g) \int_{x\geq m-g} p_{\xi|G}(x)dxdg}_{I_2}.\end{aligned}$$

For I_1, since g is negative, we can enlarge the inner domain:

$$|I_1| \leq \int_{\mathbb{R}^-} |g| p_G(g) \int_{x\geq m} p_{\xi|G}(x)dxdg = \mathbb{E}\left[|G|\mathbb{1}\{G \in \mathbb{R}^-\}\mathbb{P}(\xi > m|G)\right] \underset{m\to\infty}{=} o\left(m^{1-1/\gamma}\right)$$

by Lemma 19(i) with $W = |G|\mathbb{1}\{G \in \mathbb{R}^-\}$.

For I_2, we cut at $m + t_m$ and use that g is positive:

$$I_2 \geq \underbrace{\int_0^{m+t_m} g p_G(g) \int_{x\geq m} p_{\xi|G}(x)dxdg}_{E_1} + \underbrace{\int_{m+t_m}^{\infty} g p_G(g) \int_{x\geq m-g} p_{\xi|G}(x)dxdg}_{E_2}.$$

Lemma 19(i) applied to $W = G\mathbb{1}\{G \in [0, m + t_m)\} \leq G\mathbb{1}\{G \geq 0\}$, whose moments of order $p < 1/\gamma$ are finite by the tail assumption on G and bounded uniformly in m, gives $E_1 = o(m^{1-1/\gamma})$. For E_2:

$$E_2 \geq (m + t_m) \int_{m+t_m}^{\infty} \int_{x\geq -t_m} p_{G,\xi}(g, x)dxdg \;=\; (m + t_m)\mathbb{P}(G \geq m + t_m, \xi \geq -t_m)$$

bounding the joint probability below by the Fréchet lower bound:

$$\geq (m + t_m) \max\left(\mathbb{P}(G \geq m + t_m) - \mathbb{P}(\xi \leq -t_m), 0\right)$$

and using that, for large enough m, $\mathbb{P}(\xi \leq -t_m) \leq e^{-ct_m} = \frac{l(m)}{m^{2/\gamma}}$ by the definition of t_m:

$$\geq (m + t_m) \max\left(\frac{l(m + t_m)}{(m + t_m)^{1/\gamma}} - \frac{l(m)}{m^{2/\gamma}}, 0\right).$$

For the argument of the max:

$$\begin{aligned}\frac{l(m + t_m)}{(m + t_m)^{1/\gamma}} - \frac{l(m)}{m^{2/\gamma}} =& \frac{m^{2/\gamma} l(m + t_m) - (m + t_m)^{1/\gamma} l(m)}{(m + t_m)^{1/\gamma} m^{2/\gamma}} \\ =& \frac{l(m + t_m) - l(m)\left(\frac{m+t_m}{m^2}\right)^{1/\gamma}}{m^{1/\gamma}(1 + t_m/m)^{1/\gamma}},\end{aligned}$$

which is positive for large enough m, since $t_m = O(\log m) = o(m)$ and $\left(\frac{m+t_m}{m^2}\right)^{1/\gamma} \to 0$.

We now turn to the normalization:

$$\zeta = \underbrace{\int_{\mathbb{R}^-} p_G(g) \int_{x \geq m-g} p_{\xi|G}(x) dx dg}_{I_1'} + \underbrace{\int_0^\infty p_G(g) \int_{x \geq m-g} p_{\xi|G}(x) dx dg}_{I_2'}.$$

By Lemma 19(iii) with $W = 1$, $I_1' = o(m^{-1/\gamma})$. For I_2', let $\lambda \in (0,1)$ and cut at λm:

$$E_1' := \int_0^{\lambda m} p_G(g) \int_{x \geq m-g} p_{\xi|G}(x) dx dg \leq \mathbb{E}\left[\mathbb{1}\{G \in [0, \lambda m)\} \mathbb{P}\left(\xi > (1-\lambda)m | G\right)\right] = o\left(m^{-1/\gamma}\right),$$

$$E_2' := \int_{\lambda m}^\infty p_G(g) \int_{x \geq m-g} p_{\xi|G}(x) dx dg \leq \int_{\lambda m}^\infty p_G(g) dg = \frac{l(\lambda m)}{(\lambda m)^{1/\gamma}},$$

the first again by Lemma 19(iii), using $(1-\lambda)m \asymp m$ for fixed $\lambda < 1$.

Combining, for every $\lambda \in (0,1)$,

$$\begin{aligned}
\mathbb{E}\left[G | M > m\right] &\underset{m\to\infty}{\geq} (m + t_m) \frac{l(m+t_m) - l(m)\left(\frac{m+t_m}{m^2}\right)^{1/\gamma}}{m^{1/\gamma}(1 + t_m/m)^{1/\gamma}} \times \frac{(\lambda m)^{1/\gamma}}{l(\lambda m) + o(1)} + o(m) \\
&\underset{m\to\infty}{\sim} (m + t_m) \lambda^{1/\gamma} \frac{l(m+t_m) - l(m)\left(\frac{m+t_m}{m^2}\right)^{1/\gamma}}{l(m+t_m)} \\
&\underset{m\to\infty}{\sim} \lambda^{1/\gamma} m.
\end{aligned}$$

Since this holds for every $\lambda \in (0,1)$, $\mathbb{E}\left[G | M > m\right] \geq m + o(m)$. □

Theorem 2. *Suppose that G is heavy-tailed with tail index γ, and that the left tail of G is no heavier than its right tail, in the sense that $\mathbb{E}\left[|G|^{1/\gamma} \mathbb{1}\{G < 0\}\right] < \infty$. Suppose that the left and right tails of ξ are at most as heavy as those of a heavy-tailed variable of index α, and define $M := G + \xi$. Assume further that $1/\alpha - 1/\gamma > 0$ and $1/\gamma > 1$. Then*

$$\mathbb{E}\left[G | M > m\right] \underset{m\to\infty}{\geq} m + o(m).$$

Proof. We have

$$\begin{aligned}
&\int_{\mathbb{R}} g p_G(g) \int_{x \geq m-g} p_{\xi|G}(x) dx dg \\
=&\underbrace{\int_{\mathbb{R}^-} g p_G(g) \int_{x \geq m-g} p_{\xi|G}(x) dx dg}_{I_1} + \underbrace{\int_{\mathbb{R}^+} g p_G(g) \int_{x \geq m-g} p_{\xi|G}(x) dx dg}_{I_2}.
\end{aligned}$$

We first bound I_1:

$$\left| \int_{\mathbb{R}^-} g p_G(g) \int_{x \geq m-g} p_{\xi|G}(x) dx dg \right| \leq \int_{\mathbb{R}^-} |g| p_G(g) \int_{x \geq m-g} p_{\xi|G}(x) dx dg,$$

using the fact that g is negative:

$$\begin{aligned}
&\leq \int_{\mathbb{R}^-} |g| p_G(g) \int_{x \geq m} p_{\xi|G}(x) dx dg \\
&= \mathbb{E}\left[|G| \mathbb{1}\{G \in \mathbb{R}^-\} \mathbb{P}\left(\xi > m | G\right)\right].
\end{aligned}$$

Applying Lemma 19(ii) with $W = |G| \mathbb{1}\{G \in \mathbb{R}^-\}$, which satisfies $\mathbb{E}\left[W^{1/\gamma}\right] < \infty$ by assumption, we obtain

$$I_1 \underset{m\to\infty}{=} o\left(m^{1-1/\gamma}\right).$$

Next, we bound I_2 from below. Let t_m be a threshold, to be chosen later. Cutting the integral at $m + t_m$:

$$I_2 = \int_0^{m+t_m} g p_G(g) \int_{x \geq m-g} p_{\xi|G}(x) dx dg + \int_{m+t_m}^{\infty} g p_G(g) \int_{x \geq m-g} p_{\xi|G}(x) dx dg,$$

using this time the fact that g is positive:

$$\geq \underbrace{\int_0^{m+t_m} g p_G(g) \int_{x \geq m} p_{\xi|G}(x) dx dg}_{E_1} + \underbrace{\int_{m+t_m}^{\infty} g p_G(g) \int_{x \geq m-g} p_{\xi|G}(x) dx dg}_{E_2}.$$

E_1 is nonnegative, so we simply bound it below by 0. For E_2:

$$\begin{aligned}
E_2 \geq & \int_{m+t_m}^{\infty} g p_G(g) \int_{x \geq -t_m} p_{\xi|G}(x) dx dg \\
\geq & (m + t_m) \int_{m+t_m}^{\infty} p_G(g) \int_{x \geq -t_m} p_{\xi|G}(x) dx dg \\
= & (m + t_m) \int_{m+t_m}^{\infty} \int_{x \geq -t_m} p_{G,\xi}(g, x) dx dg \\
= & (m + t_m) \mathbb{P}\left(G \geq m + t_m, \xi \geq -t_m\right)
\end{aligned}$$

where we bound this joint probability below by the Fréchet lower bound:

$$\geq (m + t_m) max(\mathbb{P}\left(G \geq m + t_m\right) + \mathbb{P}\left(\xi \geq -t_m\right) - 1, 0).$$

Here we use the fact that $\mathbb{P}\left(\xi \leq -t_m\right) = \frac{l(t_m)}{t_m^{1/\alpha}}$, and, by Lemma 17, for every $q > 0$,

$$l(t_m) \underset{m \to +\infty}{=} o((t_m)^q),$$

which leads to

$$\mathbb{P}\left(\xi \leq -t_m\right) \underset{m \to +\infty}{=} o(t_m^{q - 1/\alpha}).$$

Here we need t_m such that $t_m = o(m)$ and $t_m^{q-1/\alpha} = o(m^{-1/\gamma - q'})$, where $q' > 0$ accounts for a possible decay of l at rate $m^{-q'}$. Writing $t_m = m^\kappa$ with $\kappa < 1$, this reads

$$\begin{aligned}
& \kappa(1/\alpha - q) > 1/\gamma + q' \\
\iff & \kappa > \frac{1/\gamma + q'}{1/\alpha - q} = \frac{\alpha}{\gamma}\left(\frac{1 + \gamma q'}{1 - \alpha q}\right).
\end{aligned} \tag{7}$$

The right-hand side is increasing in both q and q', and tends to $\frac{\alpha}{\gamma}$ as $(q, q') \to (0, 0)$. Since $1/\alpha > 1/\gamma$ is equivalent to $\alpha < \gamma$, this limit is strictly below 1; by continuity we may fix $q \in (0, 1/\alpha)$ and $q' > 0$ small enough that the right-hand side remains < 1, and then choose κ strictly between it and 1. This guarantees $t_m = o(m)$ and $\mathbb{P}\left(\xi \leq -t_m\right) = o\left(\mathbb{P}\left(G \geq m + t_m\right)\right)$. Using $t_m = m^\kappa$:

$$\begin{aligned}
& \geq \quad (m + t_m) max(\mathbb{P}\left(G \geq m + t_m\right) + \mathbb{P}\left(\xi \geq -t_m\right) - 1, 0) \\
& \underset{m \to +\infty}{\geq} (m + t_m) max(\mathbb{P}\left(G \geq m + t_m\right) - o(\frac{1}{m^{1/\gamma}}), 0) \\
& = \quad (m + t_m) max(\frac{l(m + t_m)}{(m + t_m)^{1/\gamma}} - o(\frac{1}{m^{1/\gamma}}), 0) \\
& = \quad (m + t_m) max(\frac{l(m + t_m)}{m^{1/\gamma}(1 + t_m/m)^{1/\gamma}} - o(\frac{1}{m^{1/\gamma}}), 0).
\end{aligned}$$

Considering the max:

$$\frac{l(m+t_m)}{m^{1/\gamma}(1+t_m/m)^{1/\gamma}} - o(\frac{1}{m^{1/\gamma}}) \underset{m\to\infty}{=} \frac{1}{m^{1/\gamma}}\left(\frac{l(m+t_m)}{(1+t_m/m)^{1/\gamma}} - o(1)\right). \tag{8}$$

The right-hand side of (8) is positive for large m, since $\frac{t_m}{m} \underset{m\to\infty}{\to} 0$. In the previous bound,

$$(m+t_m)max(\frac{l(m+t_m)}{(m+t_m)^{1/\gamma}} + o(\frac{1}{m^{1/\gamma}}), 0) \underset{m\to+\infty}{=} (m+t_m)\frac{1}{m^{1/\gamma}}(\frac{l(m+t_m)}{(1+t_m/m)^{1/\gamma}} + o(1)).$$

We now turn to the normalization $\mathbb{P}(M > m)$:

$$\begin{aligned}\mathbb{P}(M>m) &= \int_{\mathbb{R}} p_G(g)\int_{x\geq m-g} p_{\xi|G}(x)dxdg \\ &= \underbrace{\int_{\mathbb{R}^-} p_G(g)\int_{x\geq m-g} p_{\xi|G}(x)dxdg}_{I_1'} + \underbrace{\int_0^\infty p_G(g)\int_{x\geq m-g} p_{\xi|G}(x)dxdg}_{I_2'}.\end{aligned}$$

By the same argument as for I_1, applied with $W = 1$, which is bounded, we know that $I_1' = o\left(m^{q-1/\alpha}\right)$ for every $q > 0$; since $1/\alpha - 1/\gamma > 0$, choosing $q < 1/\alpha - 1/\gamma$ gives $I_1' = o\left(m^{-1/\gamma}\right)$. For I_2', let $\lambda \in (0,1)$:

$$I_2' = \underbrace{\int_0^{\lambda m} p_G(g)\int_{x\geq m-g} p_{\xi|G}(x)dxdg}_{E_1'} + \underbrace{\int_{\lambda m}^\infty p_G(g)\int_{x\geq m-g} p_{\xi|G}(x)dxdg}_{E_2'}.$$

For E_1':

$$\begin{aligned}\int_0^{\lambda m} p_G(g)\int_{x\geq m-g} p_{\xi|G}(x)dxdg &\leq \int_0^{\lambda m} p_G(g)\int_{x\geq(1-\lambda)m} p_{\xi|G}(x)dxdg \\ &= \mathbb{E}\left[\mathbb{1}\{G\in[0,\lambda m)\}\mathbb{P}\left(\xi > (1-\lambda)m|G\right)\right] \\ &= o\left(m^{-1/\gamma}\right),\end{aligned}$$

by Lemma 19(iii) with $W = 1$ and the fact that $(1-\lambda)m \asymp m$ for fixed $\lambda < 1$.

For E_2':

$$\begin{aligned}\int_{\lambda m}^\infty p_G(g)\int_{x\geq m-g} p_{\xi|G}(x)dxdg &\leq \int_{\lambda m}^\infty p_G(g)dg \\ &= \frac{l(\lambda m)}{(\lambda m)^{1/\gamma}}\end{aligned}$$

Therefore, for every $\lambda \in (0,1)$,

$$\begin{aligned}\mathbb{E}[G|M>m] &\underset{m\to\infty}{\geq} (m+t_m)\times\frac{(\lambda m)^{1/\gamma}}{l(\lambda m)+o(1)}\times\frac{1}{m^{1/\gamma}}\left(\frac{l(m+t_m)}{(1+t_m/m)^{1/\gamma}} - o(1)\right) + o(m) \\ &\underset{m\to\infty}{\sim} (m+t_m)\times\frac{\lambda^{1/\gamma}l(m+t_m)}{l(\lambda m)} + o(m) \\ &\underset{m\to\infty}{\sim} \lambda^{1/\gamma}m\end{aligned}$$

Since this holds for every $\lambda \in (0,1)$,

$$\mathbb{E}[G|M>m] \underset{m\to\infty}{\geq} m + o(m).$$

□