The Strong, weak and benign Goodhart's law. An independence-free and paradigm-agnostic formalisation

Adrien Majka [1] **El-Mahdi El-Mhamdi** [1]

Abstract

Goodhart's law is a famous adage in policy-making that states that "When a measure becomes a target, it ceases to be a good measure". As machine learning models and the optimisation capacity to train them grow, growing empirical evidence reinforced the belief in the validity of this law without however being formalised. Recently, a few attempts were made to formalise Goodhart's law, either by categorising variants of it, or by looking at how optimising a *proxy metric* affects the optimisation of an *intended goal*. In this work, we alleviate the simplifying independence assumption, made in previous works, and the assumption on the learning paradigm made in most of them, to study the effect of the coupling between the proxy metric and the intended goal on Goodhart's law. Our results show that in the case of light tailed goal and light tailed discrepancy, dependence does not change the nature of Goodhart's effect. However, in the light tailed goal and heavy tailed discrepancy case, we exhibit an example where over-optimisation occurs at a rate inversely proportional to the heavy tailedness of the discrepancy between the goal and the metric.

1. Introduction and related work

From Charles Goodhart's remark in the context of monetary economics (Goodhart, 1975) to its reformulation by Keith Hoskin (Hoskin, 1996) and its popularisation by Marylin Strathern(Strathern, 1997), Goodhart's law remained unformalised. It is only in recent years that a few attempts were made to transform this "law" from a popular wisdom to a well defined mathematical concept. Efforts to formalise Goodhart's law fall into 3 categories :

[1]Center for applied Mathematics of Polytechnique (CMAP), Ecole Polytechnique, Palaiseau, France. Correspondence to: Adrien Majka <adrien.majka@polytechnique.edu>.

ICML 2025 Workshop on Models of Human Feedback for AI Alignment

Formalisations on Reinforcement Learning. RL was naturally a favourite setting to formalise Goodhart-Strathern's law as the most notable cases of reward-hacking in AI appeared in reinforcement learning (RL) settings (Clark & Amodei, 2016; Amodei et al., 2016; Gao et al., 2022). The first part of (Kwa et al., 2024) shows the inefficiency of KL divergence to prevent reward hacking by showing that with heavy-tailed policy reward, a policy with arbitrarily high proxy reward but low penalty and low true reward always exists. (Skalse et al., 2022) introduces a formalisation of reward hacking in the context of policy learning in RL. It proves several results on the general existence of reward hacking policy with respect to different reward function on different sets of policies.

Formalisations on Supervised Learning. (Hennessy & Goodhart, 2020) intends to give a microfoundation to ML model to make them robust to the seminal (Lucas, 1976) critique of classic keynesian models. A regulator tries to make a prediction in a setting where, at test time, covariate can be manipulated by an agent to induce a more favorable decision from the regulator.

Paradigm-Agnostic Formalisations. Two precedent works by D.Manheim and S.Garrabrant (Manheim, 2023; Manheim & Garrabrant, 2018) provide interesting insight on general metrics design, by providing a towering view on metric potential flaw and set of case separation on Goodhart's law respectively, while not fully formalising the problem. (Zhuang & Hadfield-Menell, 2021) devises a general framework for AI overoptimisation and draw results in the case of constrained ressources and partially specified goal. Their setup is inspired by incomplete contracting, but its reliance on state-space descriptions makes it more suitable to reinforcement learning. Previous works (El-Mhamdi & Hoang, 2024; Kwa et al., 2024) provide a general formalisation of Goodhart's law, where no assumption is made on the learning paradigm and show nuanced cases where Goodhart's law holds or not depending on the relative thickness of the tail of the goal and the discrepancy. This was however done while assuming the discrepancy is independent from the goal. In our work, we follow the same paradigm-agnostic approach, but we alleviate the independence assumption.

Our key contributions are the following :

- We alleviate the independence hypothesis of previous work.
- We conduct a detailed analysis in two cases of dependence which highlight the importance of coupling.
- We propose a formal characterisation of different Goodhart's law effect that captures the different strenght of Goodhart's law.

Paper structure. The rest of this paper is organised as follows. In Section 2, we provide our formal setup, in Section 3 we first provide a comprehensive overview of our results, followed by their formal statements and sketches of proofs (detailed proofs are available in the Appendix), finally, Section 4 provides concluding remarks and discusses future work.

2. Model

We follow the same formalisation as in (El-Mhamdi & Hoang, 2024) and denote by G, M and ξ respectively the intended goal, the proxy metric being optimised, and the discrepancy between the goal and the proxy.

Namely, an agent is optimising the function M as a proxy for G and $G = M + \xi$. Optimising is modelised in a mechanism agnostic way. Indeed, conditioning on $M > m$ with $m \to \infty$ model optimisation of the proxy metric. A basic "sanity check" for whether the proxy fails to capture the goal is to see if M and G remain correlated as we optimise M. When M and G are not correlated as M is being optimised, one can suspect that optimising M does not help make progress in the intended goal G, that is the *weak Goodhart-Strathern* case introduced in the formalism of (El-Mhamdi & Hoang, 2024). When they keep being correlated, that is the *no Goodhart-Strathern* case. But looking at the correlation between G and M, as M is being optimised, is only a first check, one should focus on the intended goal's behaviour. To do so, we also evaluate the expected value of the goal, as the proxy is being optimised, i.e., $\lim_{m\to\overline{\mathfrak{S}_M}} \mathbb{E}\left[G|M > m\right]$ where for any random variable X, $\mathfrak{S}_X$ is the support of it and $\overline{\mathfrak{S}}_X = \sup \mathfrak{S}_X$, $\underline{\mathfrak{S}}_X = \inf \mathfrak{S}_X$ the superior and inferior limits of its support respectively. When $\mathbb{E}\left[G|M > m\right]$ decreases as we optimise M that is the same *strong Goodhart-Strathern* case in (El-Mhamdi & Hoang, 2024). In addition to the *weak, strong* and *no* Goodhart-Strathern, we introduce the *benign Goodhart-Strathern*. In the latter, M stops being correlated with G, but G keeps increasing as we optimise M.

3. Results

We first provide a comprehensive overview of different cases in Subsection 3.2, before giving our formal results in Subsection 3.2 and Subsection 3.3 together with proof sketches, all detailed proofs are available in the Appendix.

3.1. Overview

Precedent work supposed independence between the goal G and the discrepancy ξ in the equation defining the metric $M = G + \xi$. One of the key contributions of our work is to capture the dependency between the goal G and the discrepancy ξ in a way that enables a comprehensive analysis, and conduct the analysis in two scenarios :

First scenario (light tailed). The goal G and discrepancy ξ form a Gaussian random vector (ie they are both light tailed). Our result extends that of (El-Mhamdi & Hoang, 2024) to the case where the covariance between the goal G and the discrepancy ξ is not null. When maximising the metric (i.e conditioning on $M > m$, with $m \to \infty$), and provided that we have $\text{Var}\left(G\right) > \text{Var}\left(\xi\right)$, we have 3 results :

- The true goal G will also be maximised ($\mathbb{E}\left[G|M > m\right] \underset{m\to\infty}{\to} \infty$), although with a coefficient depending on the covariance.
- Despite the conditional expectation of G going to infinity, the correlation between the proxy metric M and the goal G goes to 0. This is an instance of what we coin the *benign Goodhart's law* (defined intuitively and formally in Table 1).
- The covariance between the goal G and the discrepancy ξ acts linearly on the correlation between the goal G and the proxy metric M when close to zero. When the same covariance is near its limiting value (ie $|\,\text{Cov}\left(G, \xi\right)| \sim \sqrt{\text{Var}\left(G\right)\text{Var}\left(\xi\right)}$), the correlation between the goal and the proxy metric can be arbitrarily close to one.

Second scenario (heavy tailed). The goal G is exponentially distributed, and the discrepancy ξ is heavy tailed with conditional law proportional to $\exp(G((x/\eta)^{b-1})x^{b-2}$. In this case we have two results :

- We subsume the findings of El-Mhamdi and Hoang (El-Mhamdi & Hoang, 2024) that a heavy tail on the discrepancy makes the conditional expectancy of the goal G goes to 0 when optimising the proxy metric M, demonstrating an instance of the strong Goodhart's law.

Table 1: Qualitative and formal definitions for each of the Goodhart's law outcomes

	Qualitative definition	Formal definition
No Goodhart	During optimisation the proxy stays informative and the goal G goes to its maximum value	$\exists m_0 \in \mathfrak{S}(M) / \forall m > m_0$, $\text{Corr}\left(G, M \vert M > m\right) > 0$ and $\mathbb{E}\left[G \vert M > m\right] \underset{m \to \overline{\mathfrak{S}}_M}{\longrightarrow} \overline{\mathfrak{S}}_G$
Benign	Despite the proxy's decreasing informativeness, the goal is maximised	$\text{Corr}\left(G, M \vert M > m\right) \underset{m \to \overline{\mathfrak{S}(M)}}{\longrightarrow} 0$ and $\mathbb{E}\left[G \vert M > m\right] \underset{m \to \overline{\mathfrak{S}(M)}}{\longrightarrow} \overline{\mathfrak{S}(G)}$
Weak	The expected value of the goal is bounded below its maximum value during optimisation	$\exists l \in S(G), l < \overline{\mathfrak{S}(G)}, \exists m_0 \in \mathfrak{S}(M)$ $/ \forall m > m_0, \mathbb{E}\left[G \vert M > m\right] < l$
Strong	The goal goes to its minimum value during the optimisation of the proxy	$\mathbb{E}\left[G \vert M > m\right] \underset{m \to \overline{\mathfrak{S}(M)}}{\longrightarrow} \underline{\mathfrak{S}(G)}$

- A bigger shape parameter for the discrepancy ξ (which imply *lighter tail*) will make the goal G goes to 0 quicker. That is, the same conditioning by $M > m$ will imply a smaller expected value for the goal with a lighter tail discrepancy.

3.2. Gaussian goal and Gaussian discrepancy

In this section, we study the double light-tailed situation where both G and ξ are Gaussian. For notation convenience, we represent G and ξ as a Gaussian random vector as follows.

$$\begin{pmatrix} G \\ \xi \end{pmatrix} \sim \mathcal{N}(0, \Sigma)$$

where $\Sigma = \begin{pmatrix} a & c \\ c & b \end{pmatrix}$ is the covariance matrix of the random vector composed by G and ξ. Here, $\text{Var}\left(G\right) = a$, $\text{Var}\left(\xi\right) = b$ and $\text{Cov}\left(G, \xi\right) = c$. The first result shows that in the Gaussian case, as long as the variance of the goal G dominates the variance of the discrepancy ξ ($\text{Var}\left(G\right) > \text{Var}\left(\xi\right)$), the goal G goes to infinity while the correlation between the goal G and the proxy metric M goes to 0. We call this situation the *benign Goodhart's law*, which is a special case of the "weak Goodhart" case introduced in (El-Mhamdi & Hoang, 2024). In this case, "benign" reflects the fact that while the correlation between the goal and the proxy is going to zero it does not prevent the goal from going to infinity.

Lemma 3.1. *With $(G, \xi) \sim \mathcal{N}(0, \Sigma)$, $M = G + \xi$ and $\text{Var}\left(G\right) > \text{Var}\left(\xi\right)$ the optimisation of the proxy metric also leads to the optimisation of the true goal.*

$$\mathbb{E}\left[G \vert M > m\right] \underset{m \to \infty}{\sim} \frac{a + c}{(a + b + 2c)} m.$$

The full proof of Lemma 3.1 is given in Appendix 5.2, bellow we provide a simple proof sketch.

Proof sketch. The proof contains two parts. The first part computes an equivalent for $\mathbb{P}\left(M > m\right)$, which is done using the equivalent for Gaussian tail :

$$\int_x^\infty e^{-u^2} du = \frac{e^{-x^2}}{2x} \sum_{n=0}^{N-1} (-1)^n \frac{(2n-1)!!}{2^n x^{2n}} + O\left(\frac{x^{-2N-1}}{\exp(x^2)}\right).$$

The second part computes the unormalised expected value from the formula of its definition, i.e.,

$$\int_{\mathbb{R}} \exp(-\sigma_2 x^2) \int_{t \geq \mu - x\theta} t \exp\left(-\sigma_1 t^2\right) dt dx.$$

□

The coefficient in front of the optimisation threshold m is most of the time < 1, except if the covariance between the goal G and the discrepancy ξ is very negative. With greater positive covariance, the discrepancy ξ will account for a greater portion of the proxy, thus decreasing the expected value of the goal G. On the contrary, with negative covariance between the goal G and ξ, an increase in the goal value induce a decrease proportional to the covariance in the discrepancy as we have $\mathbb{E}_G\left[\xi\right] = \frac{cG}{a}$. This leads to the goal being actually higher in expectancy with negative covariance, as for any level of the proxy considered, it has to compensate for the discrepancy that is negatively correlated.

We coin the term "benign" to describe the situation as despite the goal G going to infinity, the correlation between the proxy M and the goal G goes to 0

Theorem 3.1. *With $(G, \xi) \sim \mathcal{N}(0, \Sigma)$, $M = G + \xi$, the correlation between the proxy metric M and the goal G goes to zero in the limit no matter the correlation between the discrepancy ξ and the goal G.*

$$\text{Corr}\left(G, M \vert M > m\right) \underset{m \to \infty}{\sim} \frac{(a + c)\sqrt{a + b + 2c}}{m\sqrt{ab - c^2}}.$$

Table 2: Summary of results with respect to the goal G and discrepancy ξ tails in state of the art analyses and in our analysis.

	G \ ξ	Heavy tail	Light tail
Assuming independence (El-Mhamdi & Hoang, 2024; Kwa et al., 2024)	Heavy tail	Relevance of the proxy depends on the relative tail shape between G and ξ	No Goodhart
	Light tail	Weak Goodhart worsening with tail **thickness**	Benign Goodhart
No assumption on independance (**This paper**)	Heavy tail	x	x
	Light tail	**Strong** Goodhart, worsening with tail **lightness**	**Benign** Goodhart

The full proof of Theorem 3.1 is given in Appendix 5.3, bellow we provide a simple proof sketch.

Proof sketch. The proof uses the formula for conditional variance an covariance (taking X,Y and Z random variables) :

$$\begin{aligned} \mathrm{Var}\left(X|Y\right) =& \mathbb{E}\left[X^2\middle|Y\right] + \mathbb{E}\left[X|Y\right]^2, \\ \mathrm{Cov}\left(X,Y|Z\right) =& \mathbb{E}\left[XY|Z\right] + \mathbb{E}\left[X|Z\right]\mathbb{E}\left[Y|Z\right]. \end{aligned}$$

For the squared term, a computation of the conditional leads to a first tractable term, and a second one intractable. The second intractable term is the Gaussian density integrated over the half space in $\mathbb{R}^2$ delimited by $\theta x + y > \mu$:

$$\mathfrak{A} = \frac{1}{\sigma_2}\int_{\mathbb{R}}\int_{m-x\theta}^{+\infty}\exp(-\frac{\sigma_1}{2}x^2 - \frac{\sigma_2}{2}t^2)dtdx.$$

$\mathfrak{A}$ is proportionnal to $\mathbb{P}\left(X+Y>m\right)$ where X and Y are independant Gaussian random variables of variance $\frac{1}{\sigma_1}$ and $\frac{1}{\sigma_2}$. But we know that the sum of X and Y is a Gaussian random variable of variance $\frac{1}{\sigma_1} + \frac{1}{\sigma_2}$. As such, an equivalent for the tail of $Z := X+Y$ is also an equivalent for $\mathfrak{A}$. We compute it with the equivalent for Gaussian tail.

The crossed term poses no difficulty. The simple expectation is calculated as in the proof for G expectation. □

Here , we can differentiate two regimes of the covariance on the correlation equivalent, exemplified in Figure 1 :

- **When the covariance coefficient is close to 0**, moving the covariance will move almost linearly the correlation coefficient.
- **When the covariance coefficient is close to the limit**, the covariance matrix is almost degenerated and the point of the Gaussian lie very close to a line. It means $M \approx const \times G$. One of the consequence is that for any quantile we can choose, there exist a Gaussian coupling such that the correlation is arbitrarily close to 1 in that quantile.

Figure 1: Coefficient for the correlation equivalent depending on the value of the covariance

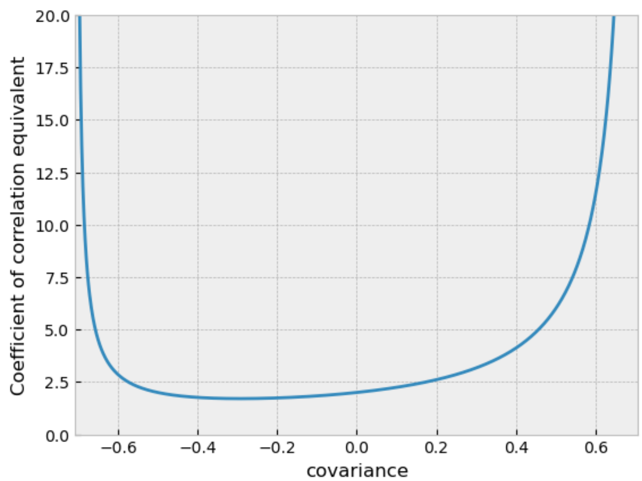

Table 3: Normal goal and discrepancy results summary

	$c<0$	$c>0$
$\mathbb{E}\left[G\middle|M>m\right]$	+	-
$\mathrm{Corr}\left(G,M\middle|M>m\right)$	-	+

The effect of covariance on conditional expectancy of the goal G and correlation between the proxy metric M and goal G is contrasted. Table 3 summarizes the covariance effect, "+" denoting an improvement on the considered quantity while "-" denotes a negative effect.

3.3. Exponential goal and heavy tailed discrepancy

In this section, we consider the case where the goal G is exponential of parameter 1. The discrepancy is then drawn conditionally to the goal G by a truncated at 1 exponential law of parameter G. This imply that ξ follow a Pareto law of shape 2 and scale 1. The same variable elevated to the power of $1/(b-1)$ and multiplied by η will be such that it follow a pareto law of shape b and scale η.

ξ then has the conditionnal density : $f_{\xi|G}(x) = G\exp(-G((\frac{x}{\eta})^{b-1}-1))\frac{x^{b-2}}{\eta^{b-1}}(b-1)\mathbb{1}\{x>\eta\}$. This case is an example of strong Goodhart's law. The optimisation the goal G tends to 0 while making the discrepancy tend to infinity.

Lemma 3.2. *When* $G \sim \mathcal{E}(1)$ *and* $f_{\xi|G}(x) = G\exp(-G((\frac{x}{\eta})^{b-1}-1))\frac{x^{b-2}}{\eta^{b-1}}(b-1)\mathbb{1}\{x>\eta\}$*, the maximisation of the proxy metric* M *also leads to the maximisation of the discrepancy* ξ*, as*

$$\mathbb{E}\left[\xi|M>m\right] \underset{m\to\infty}{\sim} m\frac{b-1}{b-2}.$$

We provides a simple proof sketch below of Lemma 3.2, the full proof can be found in Appendix 5.15

Proof sketch. The idea of the proof is inspired by Gaussian tail development. Using the fact that for any polynomial Q, we have $\frac{\mathrm{d}\exp Q(x)}{\mathrm{d}x} = Q'(x)\exp(Q(x))$, this means that any integral of the form $\int_0^{m-\eta} P(x)exp(Q(m,x))dx$ can be iteratively integrated by part to obtain as follows.

$$\begin{aligned}&\int_0^{m-\eta} P(x)exp(Q(m,x))dx\\ =&\underbrace{\sum_{n=0}^{N}\left[\frac{f^n_{Q'}(P(g))}{Q'(g)}\exp(Q(g))\right]_0^{\frac{m}{b+1}}}_{I_1}+\\ &(-1)^{N+1}\underbrace{\int_0^{m/b+1} f^{N+1}_{Q'}(P)(g)\exp(Q(g))dg}_{I_2},\end{aligned}$$

where we denote $f_{Q'}$ the operation consisting in dividing by Q' and then differentiating, $f^n_{Q'}$ consisting in applying $f_{Q'}$ n times. Using the fact that $f^{N+1}_{Q'}(P)$ is bounded (as it is a rational fraction with degree <0) by a quantity decreasing to zero with a speed depending on $N+1$, I_2 is $o(I_1)$ (using the Bachmann–Landau notations to reflect the fact that I_2 is of inferior order to I_1). The expected values are then an application of this equivalent after a some work on the original integral. □

First important thing to notice is that the discrepancy ξ is actually what's being optimised for here. The lighter tail of the goal G makes it much less likely than the discrepancy ξ to produce extreme realisation. This is true whatever the shape b of the discrepancy ξ as its tail is proportional to $1/x^b$, while the tail of the exponential goal G is proportional to $exp(-x)$. This means that, if we know that M has a very high realisation, it will be much more likely to be due to a large discrepancy ξ than a high G.

Moreover, as the discrepancy is drawn approximately following an exponential distribution of parameter G (the goal), this means that high realisation of the discrepancy ξ are associated with small realisation of the goal G. As such, the optimisation procedure - by increasing the likelihood of higher discrepancy ξ - will also make instances of very small G much more likely. This leads to following result.

Lemma 3.3. *When* $G \sim \mathcal{E}(1)$ *and* $f_{\xi|G}(x) = G\exp(-G((\frac{x}{\eta})^{b-1}-1))\frac{x^{b-2}}{\eta^{b-1}}(b-1)\mathbb{1}\{x>\eta\}$*, the goal is minimised when the proxy metric is maximised, as*

$$\mathbb{E}\left[G|M>m\right] \underset{m\to\infty}{\sim} \frac{\eta^{b-1}}{m^{b-1}}.$$

Proof sketch. See sketch of proof of Lemma 2.2 which relies on the same ideas. The full proof is available in 5.15. □

Lemma 3.3 also show that the lighter the tail of the discrepancy is (ie the bigger the shape b of ξ is), the faster the goal will decrease toward 0. This is because a light tail on the discrepancy ξ will mean more probability mass near η, which will be associated with bigger value of the goal G first. Moreover, Lemma 3.3 means that for any high value of the discrepancy ξ, it's realisation will be associated to a smaller G in expectancy if the discrepancy ξ as a lighter tail.

The two precedent results shows the importance of the coupling when talking about Goodhart's law. Indeed, if the finding here is in line with (El-Mhamdi & Hoang, 2024) with heavy tailed discrepancy, it brings nuance in the fact that the coupling here is such that the goal decrease at a speed which is actually inverse to that of the tail heaviness.

4. Concluding Remarks

Our results show that the dependence structure between the goal G and the discrepancy ξ can be of prime importance when optimising with a proxy metric M. Several natural continuation are possible :

Aggregation of metrics. In real world settings, we have access to neither the goal nor the discrepancy. However, we might have access to several proxy metrics $M_1, M_2, \ldots$ representative of the same overall goal G. Using the multiplicity of proxy metrics at disposal might be key to alleviate Goodhart's law, notably by devising aggregation rules that would make an aggregated proxy metrics $\tilde{M}$ more robust through aggregation and natural variance reduction. This would be to our sense key to alleviate alignment problem in concrete AI implementation.

Access to the proxy metric's tail. As proxy metric's tail seems to determine the presence or absence of Goodhart's law, empirical study on their prevalence within real world applications would be key to assess the importance of Goodhart's law. Devising empirically funded categories of tasks that are subject to heavy tail losses could offer a needed roadmap for practitioner to avoid reward hacking or at least be aware of potential risky situations.

Goodhart's law and evasion attacks. (Hennessy & Goodhart, 2020) study Goodhart's law within evasion attack settings[1]. So far we only considered settings were no malicious or adversarial player were present. In practical settings, AI will face adversarial behaviour that must be anticipated to avoid catastrophic failure. Creating concrete threat models as well as defense mechanisms represent a large enough scope for development.

Auditing. Auditing black box models is hard (Godinot et al., 2024). Peculiarly, developing robust and non-hackable metrics is of prime interest when auditing ML models. As such, understanding Goodhart's law can inform and strengthen research on auditing ML models.

Theoretical guarantees. Theoretical guarantees on the possibility of harmful behaviour by the AI prior or at test time is key to mitigate the global risk of AI (Bengio et al., 2024). To our sense, the line of research of this paper being by construction devised in a probabilistic setting, which is a natural way of hedging risk and assessing uncertainty on outputs, is promising. Moreover, it is sufficiently general not to constrained the scope of results to one type of algorithm or model.

[1]Often called adversarial attacks, we prefer the more specific term *evasion attacks* to distinguish them from other adversarial attacks (Vassilev et al., 2024) such as poisoning attacks and data-extraction/privacy attacks.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete Problems in AI Safety, 2016. URL https://arxiv.org/abs/1606.06565. Version Number: 2.

Bengio, Y., Cohen, M. K., Malkin, N., MacDermott, M., Fornasiere, D., Greiner, P., and Kaddar, Y. Can a Bayesian Oracle Prevent Harm from an Agent?, 2024. URL https://arxiv.org/abs/2408.05284. Version Number: 3.

Clark, J. and Amodei, D. Faulty reward functions in the wild, December 2016. URL https://openai.com/index/faulty-reward-functions/.

El-Mhamdi, E.-M. and Hoang, L.-N. On Goodhart's law, with an application to value alignment, 2024. URL https://arxiv.org/abs/2410.09638. Version Number: 1.

Gao, L., Schulman, J., and Hilton, J. Scaling Laws for Reward Model Overoptimization, 2022. URL https://arxiv.org/abs/2210.10760. Version Number: 1.

Godinot, A., Merrer, E. L., Trédan, G., Penzo, C., and Taïani, F. Under manipulations, are some AI models harder to audit? In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 644–664, Toronto, ON, Canada, April 2024. IEEE. ISBN 979-8-3503-4950-4. doi: 10.1109/SaTML59370.2024.00038. URL https://ieeexplore.ieee.org/document/10516639/.

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

Hennessy, C. and Goodhart, C. A. Goodhart's Law and Machine Learning. *SSRN Electronic Journal*, 2020. ISSN 1556-5068. doi: 10.2139/ssrn.3639508. URL https://www.ssrn.com/abstract=3639508.

Hoskin, K. 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.

Kwa, T., Thomas, D., and Garriga-Alonso, A. Catastrophic Goodhart: regularizing RLHF with KL divergence does not mitigate heavy-tailed reward misspecification, 2024. URL https://arxiv.org/abs/2407.14503. Version Number: 2.

Lucas, R. E. 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. URL https://linkinghub.elsevier.com/retrieve/pii/S0167223176800036.

Manheim, D. 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. URL https://linkinghub.elsevier.com/retrieve/pii/S2666389923002210.

Manheim, D. and Garrabrant, S. Categorizing Variants of Goodhart's Law, 2018. URL https://arxiv.org/abs/1803.04585. Version Number: 4.

Skalse, J., Howe, N. H. R., Krasheninnikov, D., and Krueger, D. Defining and Characterizing Reward Hacking, 2022. URL https://arxiv.org/abs/2209.13085. Version Number: 2.

Strathern, M. '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. URL https://www.cambridge.org/core/product/identifier/S1062798700002660/type/journal_article.

Vassilev, A., Oprea, A., Fordyce, A., and Anderson, H. Adversarial machine learning : a taxonomy and terminology of attacks and mitigations. Technical Report NIST 100-2e2023, National Institute of Standards and Technology (U.S.), Gaithersburg, MD, January 2024. URL https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf.

Zhuang, S. and Hadfield-Menell, D. Consequences of Misaligned AI. 2021. doi: 10.48550/ARXIV.2102.03896. URL https://arxiv.org/abs/2102.03896. Publisher: arXiv Version Number: 1.

5. Appendix

5.1. Normal goal and normal discrepancy :

5.1.1. Lemmas :

Lemma 5.1. $\displaystyle\int_m^{+\infty} \exp(-x^2)dx \underset{m\to\infty}{=} \frac{\exp(-m^2)}{2m}\left[1-\frac{1}{m^2}+\frac{1}{2m^4}+o(m^{-5})\right].$

Proof. This directly stems from the well known equivalent for the Gaussian tail :

$$\int_x^\infty e^{-u^2}du = \frac{e^{-x^2}}{2x}\sum_{n=0}^{N-1}(-1)^n\frac{(2n-1)!!}{2^n x^{2n}} + O(x^{-2N-1}\exp(-x^2)).$$

Applying it for $N=2$ yields the lemma □

Lemma 5.2. *For $\theta>0$, $\sigma_1>0$ and $\sigma_2>0$:*

$$\int_{\mathbb{R}} \exp\left(-\frac{x^2}{2\sigma_1}\right)\int_{t\geq m-x\theta} tx\exp\left(-\frac{t^2}{2\sigma_2}\right)dtdx = \sqrt{2\pi}\frac{2c(b+c)}{b^2}\left(\frac{ab-c^2}{a+b+2c}\right)^{3/2} m_\alpha \exp\left(-\frac{m_\alpha^2}{2(a+b+2c)}\right).$$

Proof.

$$\begin{aligned}
\mathrm{I}_1 &= \int_{\mathbb{R}} x\exp\left(-\frac{x^2}{2\sigma_1}\right)\int_{t\geq m-x\theta} t\exp\left(-\frac{t^2}{2\sigma_2}\right)dtdx \\
&= \int_{\mathbb{R}} x\exp\left(-\frac{x^2}{2\sigma_1}\right)\left[-\sigma_2\exp\left(-\frac{t^2}{2\sigma_2}\right)\right]_{m-x\theta}^{+\infty}dx \\
&= \sigma_2\int_{\mathbb{R}} x\exp\left(-\frac{x^2}{2\sigma_1}\right)\exp\left(-\frac{[m_\alpha-x\theta]^2}{2\sigma_2}\right)dx \\
&= \sigma_2\int_{\mathbb{R}} x\exp\left(-\frac{1}{2}\left(x^2\left(\frac{1}{\sigma_1}+\frac{\theta^2}{\sigma_2}\right)+\frac{m_\alpha^2}{\sigma_2}-2\frac{m_\alpha x}{\sigma_2}\right)\right)dx \\
&= \sigma_2\exp\left(-\frac{m_\alpha^2}{2\sigma_2}\right)\int_{\mathbb{R}} x\exp\left(-\frac{1}{2}\left(x^2\left(\frac{\sigma_2+\theta^2}{\sigma_2\sigma_1}\right)-2\frac{m_\alpha x}{\sigma_2}\right)\right)dx \\
&= \sigma_2\exp\left(-\frac{m_\alpha^2}{2\sigma_2}\right)\int_{\mathbb{R}} x\exp\left(-\frac{\sigma_2+\theta^2}{2\sigma_2\sigma_1}\left(x-\frac{\sigma_1}{\sigma_2+\theta^2}m_\alpha\right)^2+\frac{\sigma_1}{(\sigma_2+\theta^2)\sigma_2}m_\alpha^2\right)dx \\
&= \sigma_2\exp\left(-\frac{(\sigma_1+\sigma_2+\theta^2)m_\alpha^2}{2\sigma_2}\right)\int_{\mathbb{R}} x\exp\left(\frac{\sigma_2+\theta^2}{2\sigma_2\sigma_1}\left(x-\frac{\sigma_1}{\sigma_2+\theta^2}m_\alpha\right)^2\right)dx.
\end{aligned}$$

With the change of variable $u = x-\frac{\sigma_1}{\sigma_2+\theta^2}m_\alpha$,

$$\begin{aligned}
&= \sigma_2\exp\left(-\frac{(\sigma_1+\sigma_2+\theta^2)m_\alpha^2}{2\sigma_2}\right)\left[\underbrace{\int_{\mathbb{R}} u\exp\left(-\frac{\sigma_2+\theta^2}{2\sigma_2\sigma_1}u^2\right)du}_{=0}+\frac{\sigma_1}{\sigma_2+\theta^2}m_\alpha\int_{\mathbb{R}}\exp\left(-\frac{\sigma_2+\theta^2}{2\sigma_2\sigma_1}u^2\right)du\right] \\
&= \frac{\sigma_2\sigma_1}{\sigma_2+\theta^2}m_\alpha\exp\left(-\frac{(\sigma_1+\sigma_2+\theta^2)}{2\sigma_2\sigma_1}m_\alpha^2\right)\sqrt{\frac{2\pi\sigma_1\sigma_2}{\sigma_2+\theta^2}} \\
&= \sqrt{2\pi}\left(\frac{\sigma_2\sigma_1}{\sigma_2+\theta^2}\right)^{3/2}m_\alpha\exp\left(-\frac{(\sigma_1+\sigma_2+\theta^2)}{2\sigma_2\sigma_1}m_\alpha^2\right).
\end{aligned}$$

□

Lemma 5.3. *For $\theta > 0$, $\sigma_1 > 0$ and $\sigma_2 > 0$:*

$$\begin{aligned}
&\int_{\mathbb{R}} \exp\left(-\frac{\sigma_1}{2}x^2\right) \int_{t\geq\mu-x\theta} t^2 \exp\left(-\frac{\sigma_2}{2}t^2\right) dtdx \\
\underset{m_\alpha\to\infty}{=} &\frac{\sqrt{2\pi}}{\sqrt{\sigma_1+\theta^2\sigma_2}} \exp\left(\frac{-\mu^2\sigma_2\sigma_1}{2(\sigma_1+\theta^2\sigma_2)}\right) \\
&\times\left(\frac{\mu\sigma_1}{\sigma_2(\sigma_1+\theta^2\sigma_2)} + \frac{(\sigma_2\theta^2+\sigma_1)}{\sigma_1\sigma_2^2\mu}\left(1-\frac{(\sigma_2\theta^2+\sigma_1)}{\mu^2\sigma_1\sigma_2} + \frac{3(\sigma_2\theta^2+\sigma_1)^2}{\mu^4\sigma_1^2\sigma_2^2} + o(\mu^{-5})\right)\right).
\end{aligned}$$

Proof. First we denote :

$$\mathfrak{I} := \int_{\mathbb{R}} \exp\left(-\frac{\sigma_1}{2}x^2\right) \underbrace{\int_{t\geq\mu-x\theta} t^2 \exp\left(-\frac{\sigma_2}{2}t^2\right) dt}_{\mathfrak{I}_1}\, dx.$$

We integrate by part $\mathfrak{I}_1$:

$$\begin{aligned}
\mathfrak{I}_1 = \int_{t\geq\mu-x\theta} t^2 \exp\left(-\frac{\sigma_2}{2}t^2\right) dt &= \left[-\frac{t}{\sigma_2}\exp(-\sigma_2 t^2)\right]_{\mu-x\theta}^{\infty} + \frac{1}{\sigma_2}\int_{t\geq\mu-x\theta} \exp(\frac{\sigma_2}{2}t^2) \\
&= \frac{\mu-x\theta}{\sigma_2}\exp(\frac{\sigma_2}{2}(\mu-x\theta)^2) + \frac{1}{\sigma_2}\int_{t\geq\mu-x\theta}\exp(-\frac{\sigma_2}{2}t^2)dt.
\end{aligned}$$

Pluging it into $\mathfrak{I}$ we get :

$$\mathfrak{I} = \underbrace{\int_{\mathbb{R}} \frac{\mu-x\theta}{\sigma_2}\exp(\frac{\sigma_2}{2}(\mu^2-2\mu\theta x+\theta^2x^2) - \frac{\sigma_1}{2}x^2)dx}_{\mathfrak{I}_2} + \underbrace{\frac{1}{\sigma_2}\int_{\mathbb{R}}\int_{\mu-x\theta}^{+\infty}\exp(\frac{\sigma_1}{2}x^2 - \frac{\sigma_2}{2}t^2)dtdx}_{\mathfrak{A}}.$$

$\mathfrak{I}_2$ gives :

$$\begin{aligned}
&\int_{\mathbb{R}} \frac{\mu-x\theta}{\sigma_2}\exp\left(-\frac{\sigma_2}{2}(\mu^2-2\mu\theta x+\theta^2x^2) - \frac{\sigma_1}{2}x^2\right)dx \\
=&\int_{\mathbb{R}} \frac{\mu-x\theta}{\sigma_2}\exp\left(\frac{-\mu^2\sigma_2}{2} + 2\frac{\mu\sigma_2}{2}\theta x - (\frac{\sigma_1+\theta^2\sigma_2}{2})x^2\right)dx \\
=&\exp(\frac{-\mu^2\sigma_2}{2})\int_{\mathbb{R}} \frac{\mu-x\theta}{\sigma_2}\exp\left(-\frac{\sigma_1+\theta^2\sigma_2}{2}(x^2 - 2\frac{\mu\sigma_2}{(\sigma_1+\theta^2\sigma_2)}\theta x + \frac{\mu^2\sigma_2^2\theta^2}{(\sigma_1+\theta^2\sigma_2)^2}) + \frac{\mu^2\sigma_2^2\theta^2}{2(\sigma_1+\theta^2\sigma_2)}\right)dx \\
=&\exp\left(\frac{-\mu^2\sigma_2\left(1-\frac{\sigma_2\theta^2}{\sigma_1+\theta^2\sigma_2}\right)}{2}\right)\int_{\mathbb{R}} \frac{\mu-x\theta}{\sigma_2}\exp\left(-\frac{\sigma_1+\theta^2\sigma_2}{2}(x-\frac{\mu\sigma_2\theta}{(\sigma_1+\theta^2\sigma_2)})^2\right)dx \\
=&\exp\left(\frac{-\mu^2\sigma_2\sigma_1}{2(\sigma_1+\theta^2\sigma_2)}\right)\left(\int_{\mathbb{R}}\frac{\mu}{\sigma_2}\exp\left(-\frac{\sigma_1+\theta^2\sigma_2}{2}(x-\frac{\mu\sigma_2\theta}{(\sigma_1+\theta^2\sigma_2)})^2\right)dx\right. \\
&\left.-\int_{\mathbb{R}}\frac{x\theta}{\sigma_2}\exp\left(-\frac{\sigma_1+\theta^2\sigma_2}{2}(x-\frac{\mu\sigma_2\theta}{(\sigma_1+\theta^2\sigma_2)})^2\right)dx\right).
\end{aligned}$$

Making the change of variable $u = x - \frac{\mu\sigma_2\theta}{(\sigma_1+\theta^2\sigma_2)}$,

$$\begin{aligned}
&\int_{\mathbb{R}} \frac{\mu}{\sigma_2} \exp\left(-\frac{\sigma_1+\theta^2\sigma_2}{2}(x-\frac{\mu\sigma_2\theta}{(\sigma_1+\theta^2\sigma_2)})^2\right)dx - \int_{\mathbb{R}} \frac{x\theta}{\sigma_2} \exp\left(-\frac{\sigma_1+\theta^2\sigma_2}{2}(x-\frac{\mu\sigma_2\theta}{(\sigma_1+\theta^2\sigma_2)})^2\right)dx \\
&= \frac{\mu}{\sigma_2}\int_{\mathbb{R}} \exp\left(-\frac{\sigma_1+\theta^2\sigma_2}{2}u^2\right)dx - \int_{\mathbb{R}} \frac{(u+\frac{\mu\sigma_2\theta}{(\sigma_1+\theta^2\sigma_2)})\theta}{\sigma_2}\exp\left(-\frac{\sigma_1+\theta^2\sigma_2}{2}u^2\right)dx \\
&= \frac{\mu}{\sigma_2}\sqrt{\frac{2\pi}{\sigma_1+\theta^2\sigma_2}} - \underbrace{\int_{\mathbb{R}} \frac{u\theta}{\sigma_2}\exp\left(-\frac{\sigma_1+\theta^2\sigma_2}{2}u^2\right)dx}_{=0} - \frac{\mu\sigma_2\theta^2}{\sigma_2(\sigma_1+\theta^2\sigma_2)}\int_{\mathbb{R}}\exp\left(-\frac{\sigma_1+\theta^2\sigma_2}{2}u^2\right) \\
&= \frac{\mu}{\sigma_2}\sqrt{\frac{2\pi}{\sigma_1+\theta^2\sigma_2}} - \frac{\mu\sigma_2\theta^2}{\sigma_2(\sigma_1+\theta^2\sigma_2)}\sqrt{\frac{2\pi}{\sigma_1+\theta^2\sigma_2}}.
\end{aligned}$$

It yields for $\mathfrak{I}_2$:

$$\begin{aligned}
&\exp\left(\frac{-\mu^2\sigma_2\sigma_1}{2(\sigma_1+\theta^2\sigma_2)}\right)\sqrt{\frac{2\pi}{\sigma_1+\theta^2\sigma_2}}\left(\frac{\mu}{\sigma_2} - \frac{\mu\sigma_2\theta^2}{\sigma_2(\sigma_1+\theta^2\sigma_2)}\right) \\
=&\exp\left(\frac{-\mu^2\sigma_2\sigma_1}{2(\sigma_1+\theta^2\sigma_2)}\right)\sqrt{\frac{2\pi}{\sigma_1+\theta^2\sigma_2}}\frac{\mu\sigma_1}{\sigma_2(\sigma_1+\theta^2\sigma_2)}.
\end{aligned}$$

We want now to approximate :

$$\mathfrak{A} = \frac{1}{\sigma_2}\int_{\mathbb{R}}\int_{\mu-x\theta}^{+\infty} \exp(-\frac{\sigma_1}{2}x^2 - \frac{\sigma_2}{2}t^2)dtdx.$$

For $\mathfrak{A}$, we can remark that if $X \sim \mathrm{N}(0, \frac{\theta^2}{\sigma_1})$ and $Y \sim \mathrm{N}(0, \frac{1}{\sigma_2})$, with $X \perp\!\!\!\perp Y$ we have :

$$\mathbb{P}(X+Y \geq \mu) = \frac{\sqrt{\sigma_1\sigma_2}}{2\pi|\theta|}\int_{\mathbb{R}}\int_{\mu-y}^{+\infty}\exp\left(-x^2\frac{\sigma_1}{2\theta^2} - y^2\frac{\sigma_2}{2}\right)dxdy.$$

But if we set $u = x\theta$ in the preceding integral we get :

$$\frac{1}{|\theta|\sigma_2}\int_{\mathbb{R}}\int_{\mu-u}^{+\infty}\exp(-\frac{\sigma_1}{2\theta^2}u^2 - \frac{\sigma_2}{2}t^2)dtdu = \frac{2\pi}{\sigma_2\sqrt{\sigma_1\sigma_2}}\mathbb{P}(X+Y\geq\mu).$$

But we know that as X and Y are normal and independant, they form a Gaussian vector with diagonal variance matrix. So we can easily calculate the law of $X + Y = Z \sim \mathrm{N}(0, \frac{\sigma_2\theta^2+\sigma_1}{\sigma_1\sigma_2})$, so :

$$\frac{2\pi}{\sigma_2\sqrt{\sigma_1\sigma_2}}\mathbb{P}(X+Y\geq\mu) = \frac{2\pi}{\sigma_2\sqrt{\sigma_1\sigma_2}}\mathbb{P}(Z\geq\mu) = \frac{\sqrt{2\pi}}{\sigma_2\sqrt{(\sigma_2\theta^2+\sigma_1)}}\int_{\mu}^{+\infty}\exp\left(\frac{-z^2\sigma_1\sigma_2}{2(\sigma_2\theta^2+\sigma_1)}\right)dz.$$

By setting $t = z\sqrt{\frac{\sigma_1\sigma_2}{2(\sigma_2\theta^2+\sigma_1)}}$ we have :

$$\begin{aligned}
&\frac{1}{\sigma_2}\sqrt{\frac{2\pi}{(\sigma_2\theta^2+\sigma_1)}}\int_{\mu}^{+\infty}\exp\left(\frac{-z^2\sigma_1\sigma_2}{2(\sigma_2\theta^2+\sigma_1)}\right)dz \\
=\quad&\frac{1}{\sigma_2}\frac{2\sqrt{\pi}}{\sqrt{\sigma_1\sigma_2}}\int_{\sqrt{\frac{\sigma_1\sigma_2}{2(\sigma_2\theta^2+\sigma_1)}}\mu}^{+\infty}\exp\left(-t^2\right)dt.
\end{aligned}$$

Using here the lemma 1 with $N = 3$

$$\underset{\mu\to\infty}{=}\frac{1}{\sigma_2}\frac{\sqrt{2\pi(\sigma_2\theta^2+\sigma_1)}}{\sigma_1\sigma_2\mu}\exp\left(\frac{-\mu^2\sigma_1\sigma_2}{2(\sigma_2\theta^2+\sigma_1)}\right)\left(1 - \frac{(\sigma_2\theta^2+\sigma_1)}{\mu^2\sigma_1\sigma_2} + \frac{2(\sigma_2\theta^2+\sigma_1)^2}{\mu^4\sigma_1^2\sigma_2^2} + o(\mu^{-5})\right),$$

hence the result. □

Lemma 5.4. *For $\sigma_1 > 0$ and $\sigma_2 > 0$:*

$$\int_{\mathbb{R}} \exp(-\sigma_2 x^2) \int_{t\geq \mu - x\theta} t \exp\left(-\sigma_1 t^2\right) dt dx = \frac{\sqrt{\pi}}{2\sigma_1\sqrt{\sigma_2+\sigma_1\theta^2}} \exp\left(-\mu^2 \frac{\sigma_1\sigma_2}{\sigma_2+\sigma_1\theta^2}\right).$$

Proof. Setting

$$\mathfrak{L} := \int_{\mathbb{R}} \exp(-\sigma_2 x^2) \int_{t\geq \mu - x\theta} t \exp\left(-\sigma_1 t^2\right) dt dx,$$

we have :

$$\begin{aligned}
\mathfrak{L} &= \int_{\mathbb{R}} \exp(-\sigma_2 x^2) \left[-\frac{\exp(-\sigma_1 t^2)}{2\sigma_1}\right]_{\mu - x\theta}^{+\infty} dx \\
&= \frac{1}{2\sigma_1} \int_{\mathbb{R}} \exp(-\sigma_2 x^2 - \sigma_1(\mu - x\theta)^2) dx \\
&= \frac{1}{2\sigma_1} \int_{\mathbb{R}} \exp(-\sigma_2 x^2 - \sigma_1\mu^2 - \sigma_1 x^2\theta^2 + 2\sigma_1\theta\mu x) dx \\
&= \frac{1}{2\sigma_1} \exp(-\sigma_1\mu^2) \int_{\mathbb{R}} \exp(-x^2(\sigma_2 + \sigma_1\theta^2) + 2\sigma_1\theta\mu x) dx \\
&= \frac{1}{2\sigma_1} \exp(-\sigma_1\mu^2) \int_{\mathbb{R}} \exp\left(-(\sigma_2 + \sigma_1\theta^2)(x^2 + 2\frac{\sigma_1\theta}{\sigma_2+\sigma_1\theta^2}\mu x)\right) dx \\
&= \frac{1}{2\sigma_1} \exp\left(-\mu^2(\sigma_1 - \frac{\sigma_1^2\theta^2}{\sigma_2+\sigma_1\theta^2})\right) \int_{\mathbb{R}} \exp\left(-(\sigma_2 + \sigma_1\theta^2)(x + \frac{\sigma_1\theta}{\sigma_2+\sigma_1\theta^2}\mu)^2\right) dx \\
&= \frac{1}{2\sigma_1} \exp\left(-\mu^2 \frac{\sigma_1\sigma_2}{\sigma_2+\sigma_1\theta^2}\right) \int_{\mathbb{R}} \exp\left(-(\sigma_2 + \sigma_1\theta^2)(x + \frac{\sigma_1\theta}{\sigma_2+\sigma_1\theta^2}\mu)^2\right) dx \\
&= \frac{\sqrt{\pi}}{2\sigma_1\sqrt{\sigma_2+\sigma_1\theta^2}} \exp\left(-\mu^2 \frac{\sigma_1\sigma_2}{\sigma_2+\sigma_1\theta^2}\right).
\end{aligned}$$

□

5.2. Proof of Lemma 3.1

Lemma 3.1. *We set $M = G + \xi$ where $\begin{bmatrix} G \\ \xi \end{bmatrix} \sim \mathcal{N}(0_2, \Sigma)$, $\Sigma = \begin{bmatrix} a & c \\ c & b \end{bmatrix}$ with $a > 0$, $b > 0$ and $|c| < \sqrt{ab}$. Then :*

$$\mathbb{P}\left(M > m_\alpha\right) \underset{m_\alpha \to \infty}{=} \frac{\sqrt{a+b+2c}}{\sqrt{2\pi}m_\alpha} \exp\left(\frac{-m_\alpha^2}{2(a+b+2c)}\right) \left[1 - \frac{a+b+2c}{m_\alpha^2} + \frac{3(a+b+2c)^2}{m_\alpha^4} + o(m_\alpha^{-5})\right], \tag{1}$$

$$\mathbb{E}\left[G|M > m_\alpha\right] \underset{m_\alpha \to \infty}{=} \frac{a+c}{(a+b+2c)} m_\alpha (1 + \frac{a+b+2c}{m_\alpha^2} - 2\frac{(a+b+2c)^2}{m_\alpha^4} + o(m_\alpha^{-5})), \tag{2}$$

$$\mathbb{E}\left[\xi|M > m_\alpha\right] \underset{m_\alpha \to \infty}{=} \frac{(b+c)}{(a+b+2c)} m_\alpha (1 + \frac{a+b+2c}{m_\alpha^2} - 2\frac{(a+b+2c)^2}{m_\alpha^4} + o(m_\alpha^{-5})), \tag{3}$$

$$\mathbb{E}\left[G\xi|M > m_\alpha\right] \underset{m_\alpha \to \infty}{=} \frac{(a+c)(b+c)}{(a+b+2c)^2} m_\alpha^2 + \frac{(a+c)(b+c)}{(a+b+2c)} + c - 2\frac{(a+c)(b+c)}{m_\alpha^2} + o(m_\alpha^{-3}), \tag{4}$$

$$\mathbb{E}\left[G^2\middle|M > m_\alpha\right] \underset{m_\alpha \to \infty}{=} \frac{(a+c)^2}{(a+b+2c)^2} m_\alpha^2 + \frac{(a+c)^2}{(a+b+2c)} + a - 2\frac{(a+c)^2}{m_\alpha^2} + o(m_\alpha^{-3}). \tag{5}$$

Proof. **For (1):** As $\begin{bmatrix} G \\ \xi \end{bmatrix} \sim \mathcal{N}(0_2, \Sigma)$, we have $M \sim \mathrm{N}(0, a+b+2c)$:

$$\mathbb{P}\left(M \geq m_\alpha\right) = \frac{1}{\sqrt{2\pi(a+b+2c)}} \int_{m_\alpha}^{+\infty} \exp(\frac{-x^2}{2(a+b+2c)}) dx.$$

With $u = \frac{x}{\sqrt{2(a+b+2c}}$:

$$= \frac{1}{\sqrt{\pi}} \int_{\frac{m_\alpha}{\sqrt{2(a+b+2c)}}}^{+\infty} \exp(-u^2) du.$$

Using here the lemma 1 for $N = 3$

$$\underset{m_\alpha \to \infty}{=} \frac{\sqrt{a+b+2c}}{\sqrt{2\pi} m_\alpha} \exp\left(\frac{-m_\alpha^2}{2(a+b+2c)}\right) \left[1 - \frac{a+b+2c}{m_\alpha^2} + \frac{3(a+b+2c)^2}{m_\alpha^4} + o(m_\alpha^{-5})\right].$$

For (2):

$$\begin{aligned}
\mathbb{E}\left[G | G + \xi \geq m_\alpha\right] &= \frac{1}{\alpha 2\pi \sqrt{ab - c^2}} \int_{\mathbb{R}} \int_{g+x \geq m_\alpha} g \exp\left(-\frac{\delta}{2}(bg^2 - 2cgx + ax^2)\right) dg dx \\
&= \frac{1}{\alpha 2\pi \sqrt{ab - c^2}} \int_{\mathbb{R}} \exp\left(-\frac{x^2}{2b}\right) \int_{g+x \geq m_\alpha} g \exp\left(-\frac{\delta b}{2}(g - \frac{c}{b}x)^2\right) dg dx,
\end{aligned}$$

with $t = g - \frac{c}{b}x$:

$$= \frac{1}{\alpha 2\pi \sqrt{ab - c^2}} \int_{\mathbb{R}} \exp\left(-\frac{x^2}{2b}\right) \int_{t \geq m - x(1+\frac{c}{b})} (t + \frac{c}{b}x) \exp\left(-\frac{\delta b}{2} t^2\right) dt dx.$$

Splitting it in two :

$$\begin{aligned}
\mathrm{A}_1 &:= \frac{1}{\alpha 2\pi \sqrt{ab - c^2}} \int_{\mathbb{R}} \int_{t \geq m - x(1+\frac{c}{b})} t \exp\left(-\frac{\delta b}{2} t^2 - \frac{x^2}{2b}\right) dt dx \\
\mathrm{A}_2 &:= \frac{1}{\alpha 2\pi \sqrt{ab - c^2}} \frac{c}{b} \int_{\mathbb{R}} \int_{t \geq m - x(1+\frac{c}{b})} x \exp\left(-\frac{\delta b}{2} t^2 - \frac{x^2}{2b}\right) dt dx.
\end{aligned}$$

We can use here Lemma 5.4 two times, which gives :

$$\alpha \mathrm{A}_1 = \frac{\sqrt{2\pi}(ab - c^2)^{3/2}}{b\sqrt{a+b+2c}}.$$

And with normalisation :

$$\mathrm{A}_1 \underset{m_\alpha \to +\infty}{=} \frac{(ab - c^2) m_\alpha}{b(a+b+2c)} (1 + \frac{a+b+2c}{m_\alpha^2} - 2\frac{(a+b+2c)^2}{m_\alpha^4} + o(m_\alpha^{-5})).$$

In the same way :

$$\alpha \mathrm{A}_2 = \frac{(b+c)c}{b\sqrt{2\pi(b+2c+a)}} \exp\left(-m_\alpha^2 \frac{1}{2(a+b+2c)}\right).$$

And after normalisation:

$$\mathrm{A}_2 \underset{m_\alpha \to \infty}{=} \frac{(b+c)c}{b(a+b+2c)} m_\alpha (1 + \frac{a+b+2c}{m_\alpha^2} - 2\frac{(a+b+2c)^2}{m_\alpha^4} + o(m_\alpha^{-5})),$$

hence the result.

For (3):

$$\begin{aligned}\mathbb{E}\left[\xi|G+\xi \geq m_\alpha\right] &= \frac{1}{\alpha 2\pi\sqrt{ab-c^2}}\int_{\mathbb{R}}\int_{g+x\geq m_\alpha} x\exp\left(-\frac{\delta}{2}(bg^2-2cgx+ax^2)\right)dgdx \\ &= \frac{1}{\alpha 2\pi\sqrt{ab-c^2}}\int_{\mathbb{R}}\exp\left(-\frac{x^2}{2b}\right)\int_{g+x\geq m_\alpha} x\exp\left(-\frac{\delta b}{2}(g-\frac{c}{b}x)^2\right).\end{aligned}$$

With $t = g - \frac{c}{b}$:

$$\begin{aligned}&= \frac{1}{\alpha 2\pi\sqrt{ab-c^2}}\int_{\mathbb{R}}\exp\left(-\frac{x^2}{2b}\right)\int_{t\geq m-x(1+\frac{c}{b})} x\exp\left(-\frac{\delta b}{2}t^2\right)dtdx \\ &= \frac{b}{c}\mathrm{A}_2.\end{aligned}$$

So :

$$\mathbb{E}\left[\xi|G+\xi \geq m_\alpha\right] \underset{m_\alpha\to\infty}{=} \frac{(b+c)}{(a+b+2c)}m_\alpha(1+\frac{a+b+2c}{m_\alpha^2}-2\frac{(a+b+2c)^2}{m_\alpha^4}+o(m_\alpha^{-5})).$$

For (4):

$$\begin{aligned}\mathbb{E}_\alpha\left[G\xi\right] &= \frac{1}{\alpha 2\pi\sqrt{ab-c^2}}\int_{\mathbb{R}}\int_{g+x\geq m_\alpha} gx\exp\left(-\frac{\delta}{2}(bg^2-2cgx+ax^2)\right)dgdx \\ &= \frac{1}{\alpha 2\pi\sqrt{ab-c^2}}\int_{\mathbb{R}}x\exp\left(-\frac{x^2}{2b}\right)\int_{g+x\geq m_\alpha} g\exp\left(-\frac{\delta b}{2}(g-\frac{c}{b}x)^2\right)dgdx.\end{aligned}$$

With $t = g - \frac{c}{b}x$:

$$= \frac{1}{\alpha 2\pi\sqrt{ab-c^2}}\int_{\mathbb{R}}x\exp\left(-\frac{x^2}{2b}\right)\int_{t\geq m-x(1+\frac{c}{b})}(t+\frac{c}{b}x)\exp\left(-\frac{\delta b}{2}t^2\right)dtdx.$$

We can divide it in two :

$$\begin{aligned}&\int_{\mathbb{R}}x\exp\left(-\frac{x^2}{2b}\right)\int_{t\geq m-x(1+\frac{c}{b})} t\exp\left(-\frac{\delta b}{2}t^2\right)dtdx, \\ &\frac{c}{b}\int_{\mathbb{R}}x^2\exp\left(-\frac{x^2}{2b}\right)\int_{t\geq m-x(1+\frac{c}{b})}\exp\left(-\frac{\delta b}{2}t^2\right)dtdx.\end{aligned}$$

With application of Lemma 5.3 for the first integral and Lemma 5.2 for the second we get :

$$\begin{aligned}\mathbb{E}\left[G\xi|M > m\alpha\right] &\underset{m_\alpha\to\infty}{=} \frac{(b+c)(ab-c^2)}{b(a+b+2c)^2}\left(m_\alpha^2+(a+b+2c)-\frac{2(a+b+2c)^2}{m_\alpha^2}+o(m_\alpha^{-5})\right)+ \\ &\quad \frac{c(b+c)^2m_\alpha^2}{b(a+b+2c)^2}+\frac{c(b+c)^2}{b(a+b+2c)}-2\frac{c(b+c)^2}{bm_\alpha^2}+c+o(m_\alpha^{-3}) \\ &\underset{m_\alpha\to\infty}{=} \frac{(a+c)(b+c)}{(a+b+2c)^2}m_\alpha^2+\frac{(a+c)(b+c)}{(a+b+2c)}+c-2\frac{(a+c)(b+c)}{m_\alpha^2}+o(m_\alpha^{-3}).\end{aligned}$$

For (5) :

$$\begin{aligned}
\mathbb{E}_\alpha\left[G^2\right] &= \frac{1}{\alpha 2\pi\sqrt{ab-c^2}}\int_{\mathbb{R}}\int_{g+x\geq m_\alpha} g^2 \exp\left(-\frac{\delta}{2}(bg^2-2cgx+ax^2)\right)dgdx \\
&= \frac{1}{\alpha 2\pi\sqrt{ab-c^2}}\int_{\mathbb{R}}\int_{g+x\geq m_\alpha} g^2 \exp\left(-\frac{\delta}{2}\left[b(g^2-2\frac{c}{b}gx)+ax^2\right]\right)dgdx \\
&= \frac{1}{\alpha 2\pi\sqrt{ab-c^2}}\int_{\mathbb{R}}\int_{g+x\geq m_\alpha} g^2 \exp\left(-\frac{\delta}{2}\left[b(g^2-2\frac{c}{b}gx+\frac{c^2}{b^2}x^2)-\frac{c^2}{b}x^2+ax^2]\right]\right)dgdx \\
&= \frac{1}{\alpha 2\pi\sqrt{ab-c^2}}\int_{\mathbb{R}}\int_{g+x\geq m_\alpha} g^2 \exp\left(-\frac{\delta b}{2}(g-\frac{c}{b}x)^2\right)\exp\left(-\frac{\delta}{2}\left[ax^2-\frac{c^2}{b}x^2\right]\right)dgdx \\
&= \frac{1}{\alpha 2\pi\sqrt{ab-c^2}}\int_{\mathbb{R}}\exp\left(-\frac{x^2}{2b}\right)\int_{g+x\geq m_\alpha} g^2 \exp\left(-\frac{\delta b}{2}(g-\frac{c}{b}x)^2\right)dgdx.
\end{aligned}$$

We make the following changes of variables in the second integral :

$$t = g - \frac{c}{b}x,$$

which give the bound : $t \geq m - x(1+\frac{c}{b})$
.

$$= \frac{1}{\alpha 2\pi\sqrt{ab-c^2}}\int_{\mathbb{R}}\exp\left(-\frac{x^2}{2b}\right)\int_{t\geq m-x(1+\frac{c}{b})}(t+\frac{c}{b}x)^2\exp\left(-\frac{\delta b}{2}t^2\right)dtdx.$$

We will treat the precedent integral by decomposing into 3 pieces :

$$\begin{aligned}
\mathrm{I}_1 &= \int_{\mathbb{R}}\exp\left(-\frac{x^2}{2b}\right)\int_{t\geq m-x(1+\frac{c}{b})} 2\frac{c}{b}tx\exp\left(-\frac{\delta b}{2}t^2\right)dtdx, \\
\mathrm{I}_2 &= \int_{\mathbb{R}}\exp\left(-\frac{x^2}{2b}\right)\int_{t\geq m-x(1+\frac{c}{b})} x^2\frac{c^2}{b^2}\exp\left(-\frac{\delta b}{2}t^2\right)dtdx, \\
\mathrm{I}_3 &= \int_{\mathbb{R}}\exp\left(-\frac{x^2}{2b}\right)\int_{t\geq m-x(1+\frac{c}{b})} t^2\exp\left(-\frac{\delta b}{2}t^2\right)dtdx.
\end{aligned}$$

Using Lemma 5.2 and Lemma 5.3 we obtain :

$$\mathbb{E}\left[G^2|M>m_\alpha\right] \underset{m_\alpha\to\infty}{=} \frac{(a+c)^2}{(a+b+2c)^2}m_\alpha^2 + \frac{(a+c)^2}{(a+b+2c)} + a - 2\frac{(a+c)^2}{m_\alpha^2}.$$

□

Lemma 5.5.

$$\mathrm{Var}\left(G|M>m_\alpha\right) \underset{m_\alpha\to\infty}{=} \frac{ab-c^2}{(a+b+2c)} + \frac{(a+c)^2}{m_\alpha^2} + o(m_\alpha^{-3})), \tag{6}$$

$$\mathrm{Var}\left(\xi|M>m_\alpha\right) \underset{m_\alpha\to\infty}{=} \frac{ab-c^2}{(a+b+2c)} + \frac{(b+c)^2}{m_\alpha^2} + o(m_\alpha^{-3})), \tag{7}$$

$$\mathrm{Cov}\left(G,M|M>m_\alpha\right) \underset{m_\alpha\to\infty}{=} \frac{(a+c)(a+b+2c)}{m_\alpha^2} + o(m_\alpha^{-3}). \tag{8}$$

Proof. **for 7 and 8** we use that for any random variable X and Y

$$\mathrm{Var}\left(X|Y\right) = \mathbb{E}\left[X^2\middle|Y\right] - \mathbb{E}\left[X|Y\right]^2.$$

The results then follow thanks to Section 5.2 results.

For (9), it's only a combination of what we have done precedently

$$\begin{aligned}
\mathrm{Cov}_\alpha\left(G, M\right) &= \mathbb{E}_\alpha\left[G^2\right] + \mathbb{E}_\alpha\left[G\xi\right] - \mathbb{E}_\alpha\left[G\right]^2 - \mathbb{E}_\alpha\left[G\right]\mathbb{E}_\alpha\left[\xi\right] \\
&\underset{m_\alpha\to\infty}{=} \frac{(a+c)^2}{(a+b+2c)^2}m_\alpha^2 + \frac{(a+c)^2}{(a+b+2c)} + a - 2\frac{(a+c)^2}{m_\alpha^2} \\
&\quad + \frac{(a+c)(b+c)}{(a+b+2c)^2}m_\alpha^2 + \frac{(a+c)(b+c)}{(a+b+2c)} + c - 2\frac{(a+c)(b+c)}{m_\alpha^2} + o(m_\alpha^{-3}) \\
&\quad - (\frac{a+c}{(a+b+2c)}m_\alpha(1 + \frac{a+b+2c}{m_\alpha^2} - 2\frac{(a+b+2c)^2}{m_\alpha^4} + o(m_\alpha^{-5})))^2 \\
&\quad - \frac{(b+c)}{(a+b+2c)}m_\alpha(1 + \frac{a+b+2c}{m_\alpha^2} - 2\frac{(a+b+2c)^2}{m_\alpha^4} + o(m_\alpha^{-5})) \\
&\quad \times \frac{a+c}{(a+b+2c)}m_\alpha(1 + \frac{a+b+2c}{m_\alpha^2} - 2\frac{(a+b+2c)^2}{m_\alpha^4} + o(m_\alpha^{-5})) \\
&\underset{m_\alpha\to\infty}{=} \frac{(a+c)^2}{(a+b+2c)^2}m_\alpha^2 + \frac{(a+c)^2}{(a+b+2c)} + a - 2\frac{(a+c)^2}{m_\alpha^2} \\
&\quad + \frac{(a+c)(b+c)}{(a+b+2c)^2}m_\alpha^2 + \frac{(a+c)(b+c)}{(a+b+2c)} + c - 2\frac{(a+c)(b+c)}{m_\alpha^2} + o(m_\alpha^{-3}) \\
&\quad - \frac{(a+c)^2}{(a+b+2c)^2}m_\alpha^2(1 + \frac{a+b+2c}{m_\alpha^2} - 2\frac{(a+b+2c)^2}{m_\alpha^4} + o(m_\alpha^{-5}))^2 \\
&\quad - \frac{(b+c)(a+c)}{(a+b+2c)^2}m_\alpha^2(1 + \frac{a+b+2c}{m_\alpha^2} - 2\frac{(a+b+2c)^2}{m_\alpha^4} + o(m_\alpha^{-5}))^2 \\
&\underset{m_\alpha\to\infty}{=} \frac{(a+c)^2}{(a+b+2c)^2}m_\alpha^2 + \frac{(a+c)^2}{(a+b+2c)} + a - 2\frac{(a+c)^2}{m_\alpha^2} \\
&\quad + \frac{(a+c)(b+c)}{(a+b+2c)^2}m_\alpha^2 + \frac{(a+c)(b+c)}{(a+b+2c)} + c - 2\frac{(a+c)(b+c)}{m_\alpha^2} + o(m_\alpha^{-3}) \\
&\quad - \frac{(a+c)^2}{(a+b+2c)^2}m_\alpha^2(1 + 2\frac{a+b+2c}{m_\alpha^2} - 3\frac{(a+b+2c)^2}{m_\alpha^4} + o(m_\alpha^{-5})) \\
&\quad - \frac{(b+c)(a+c)}{(a+b+2c)^2}m_\alpha^2(1 + 2\frac{a+b+2c}{m_\alpha^2} - 3\frac{(a+b+2c)^2}{m_\alpha^4} + o(m_\alpha^{-5})) \\
&\underset{m_\alpha\to\infty}{=} a + c - \frac{(a+c)^2}{(a+b+2c)} - \frac{(b+c)(a+c)}{(a+b+2c)} + \frac{(b+c)(a+c)}{m_\alpha^2} + \frac{(a+c)^2}{m_\alpha^2} + o(m_\alpha^{-3}) \\
&\underset{m_\alpha\to\infty}{=} \frac{(b+c)(a+c)}{m_\alpha^2} + \frac{(a+c)^2}{m_\alpha^2} + o(m_\alpha^{-3}) \\
&\underset{m_\alpha\to\infty}{=} \frac{(a+c)(a+b+2c)}{m_\alpha^2} + o(m_\alpha^{-3}).
\end{aligned}$$

□

5.3. Proof of Theorem 3.1

Theorem 3.1. *With $(G, \xi) \sim \mathcal{N}(0, \Sigma)$, $M = G + \xi$, the correlation between the proxy metric M and the goal G goes to zero in the limit no matter the correlation between the discrepancy ξ and the goal G.*

$$\operatorname{Corr}(G, M | M > m) \underset{m \to \infty}{\sim} \frac{(a+c)\sqrt{a+b+2c}}{m\sqrt{ab-c^2}}.$$

Proof. We have :

$$\rho_\alpha := \frac{\operatorname{Cov}_\alpha(M, G)}{\sqrt{\operatorname{Var}_\alpha(M)\operatorname{Var}_\alpha(G)}},$$

and :

$$\operatorname{Cov}_\alpha(M, G) \quad = \quad \mathbb{E}_\alpha\left[G^2\right] + \mathbb{E}_\alpha[G\xi] - \mathbb{E}_\alpha[G]^2 - \mathbb{E}_\alpha[G]\,\mathbb{E}_\alpha[\xi].$$

With the Lemma 5.5:

$$\begin{aligned}
&\underset{m_\alpha \to \infty}{=} \frac{(a+c)^2}{(a+b+2c)^2} m_\alpha^2 + \frac{(a+c)^2}{(a+b+2c)} + a - 2\frac{(a+c)^2}{m_\alpha^2} \\
&\quad + \frac{(a+c)(b+c)}{(a+b+2c)^2} m_\alpha^2 + \frac{(a+c)(b+c)}{(a+b+2c)} + c - 2\frac{(a+c)(b+c)}{m_\alpha^2} + o(m_\alpha^{-3}) \\
&\quad - (\frac{a+c}{(a+b+2c)} m_\alpha (1 + \frac{a+b+2c}{m_\alpha^2} - 2\frac{(a+b+2c)^2}{m_\alpha^4} + o(m_\alpha^{-5})))^2 \\
&\quad - \frac{(b+c)}{(a+b+2c)} m_\alpha (1 + \frac{a+b+2c}{m_\alpha^2} - 2\frac{(a+b+2c)^2}{m_\alpha^4} + o(m_\alpha^{-5})) \\
&\quad \times \frac{a+c}{(a+b+2c)} m_\alpha (1 + \frac{a+b+2c}{m_\alpha^2} - 2\frac{(a+b+2c)^2}{m_\alpha^4} + o(m_\alpha^{-5})),
\end{aligned}$$

after simplification :

$$\underset{m_\alpha \to \infty}{=} \frac{(a+c)(a+b+2c)}{m_\alpha^2} + o(m_\alpha^{-3}).$$

Then we have for the denominator :

$$\begin{aligned}
\sqrt{\operatorname{Var}_\alpha(G)\operatorname{Var}_\alpha(M)} &= \sqrt{\operatorname{Var}_\alpha(G)\operatorname{Var}_\alpha(G+\xi)} \\
&= \sqrt{\operatorname{Var}_\alpha(G)\left(\operatorname{Var}_\alpha(G) + \operatorname{Var}_\alpha(\xi) + 2\operatorname{Cov}_\alpha(G, \xi)\right)}.
\end{aligned}$$

For the covariance of ξ and G we can use Section 5.2,

$$\operatorname{Cov}_\alpha(G, \xi) \quad = \quad \mathbb{E}_\alpha[G\xi] - \mathbb{E}_\alpha[G]\,\mathbb{E}_\alpha[\xi]$$

$$\begin{aligned}
&\underset{m_\alpha \to +\infty}{=} \frac{(a+c)(b+c)}{(a+b+2c)^2} m_\alpha^2 + \frac{(a+c)(b+c)}{(a+b+2c)} + c - 2\frac{(a+c)(b+c)}{m_\alpha^2} + o(m_\alpha^{-3}) \\
&\quad - \frac{(b+c)(a+c)}{(a+b+2c)^2} m_\alpha^2 (1 + 2\frac{a+b+2c}{m_\alpha^2} - 3\frac{(a+b+2c)^2}{m_\alpha^4} + o(m_\alpha^{-5}))^2 \\
&\underset{m_\alpha \to +\infty}{=} c - \frac{(a+c)(b+c)}{(a+b+2c)} + \frac{(a+c)(b+c)}{m_\alpha^2} + o(m_\alpha^{-3}) \\
&\underset{m_\alpha \to +\infty}{=} c - \frac{(a+c)(b+c)}{(a+b+2c)} + \frac{(a+c)(b+c)}{m_\alpha^2} + o(m_\alpha^{-3}) \\
&\underset{m_\alpha \to +\infty}{=} \frac{c^2-ab}{(a+b+2c)} + \frac{(a+c)(b+c)}{m_\alpha^2} + o(m_\alpha^{-3}).
\end{aligned}$$

Using Lemma 5.5 for the variance, we have :

$$\begin{aligned}\mathrm{Var}_\alpha(G) + \mathrm{Var}_\alpha(\xi) + 2\,\mathrm{Cov}_\alpha(G,\xi) &\underset{m_\alpha\to\infty}{=} \frac{ab-c^2}{(a+b+2c)} + \frac{(a+c)^2}{m_\alpha^2} + \frac{ab-c^2}{(a+b+2c)} + \frac{(b+c)^2}{m_\alpha^2} \\ &\quad + 2\frac{c^2-ab}{(a+b+2c)} + 2\frac{(a+c)(b+c)}{m_\alpha^2} + o(m_\alpha^{-3}) \\ &\underset{m_\alpha\to\infty}{=} \frac{(a+c)^2}{m_\alpha^2} + \frac{(b+c)^2}{m_\alpha^2} + 2\frac{(a+c)(b+c)}{m_\alpha^2} + o(m_\alpha^{-3}).\end{aligned}$$

Then :

$$\begin{aligned}\mathrm{Var}_\alpha(G)\,\mathrm{Var}_\alpha(M) &\underset{m_\alpha\to\infty}{=} \left(\frac{ab-c^2}{(a+b+2c)} + \frac{(a+c)^2}{m_\alpha^2} + o(m_\alpha^{-3}))\right) \times \left(\frac{(a+c)^2}{m_\alpha^2} + \frac{(b+c)^2}{m_\alpha^2} + 2\frac{(a+c)(b+c)}{m_\alpha^2} + o(m_\alpha^{-3})\right) \\ &\underset{m_\alpha\to\infty}{=} \frac{ab-c^2}{a+b+2c}\left(\frac{(a+c)^2}{m_\alpha^2} + \frac{(b+c)^2}{m_\alpha^2} + 2\frac{(a+c)(b+c)}{m_\alpha^2} + o(m_\alpha^{-3})\right) \\ &\underset{m_\alpha\to\infty}{=} \frac{ab-c^2}{a+b+2c}\frac{(a+b+2c)^2}{m_\alpha^2} \\ &\underset{m_\alpha\to\infty}{=} \frac{(ab-c^2)(a+b+2c)}{m_\alpha^2} + o(m_\alpha^{-3}).\end{aligned}$$

Hence :

$$\sqrt{\mathrm{Var}_\alpha(G)\,\mathrm{Var}_\alpha(M)} \underset{m_\alpha\to\infty}{\sim} \frac{\sqrt{(ab-c^2)(a+b+2c)}}{m_\alpha}.$$

This finally gives :

$$\begin{aligned}\rho_\alpha &\underset{m_\alpha\to+\infty}{\sim} \frac{m_\alpha}{\sqrt{(ab-c^2)(a+b+2c)}}\frac{(a+c)(a+b+2c)}{m_\alpha^2} \\ &\underset{m_\alpha\to+\infty}{\sim} \frac{(a+c)\sqrt{a+b+2c}}{m_\alpha\sqrt{ab-c^2}}.\end{aligned}$$

□

5.4. Exponential goal and heavy tail discrepancy :

For this case, we set the goal to have an exponential law. The conditional law of the discrepancy knowing the goal is of the form (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 defined like this follow a power law of shape parameter b and position parameter η.

5.5. Lemmas :

We need the two following lemma that will be useful in near all of our next demonstration :

Lemma 5.6. *Let's consider Q a polynomial and P a rational polynomial over an interval I with $\forall x \in I, Q(x) \neq 0$ and, $\exists K \in \mathbb{R} \forall x \in I, |P(x)| \leq K$. We denote f_Q the operation such that $f_Q(P) = \frac{\partial \frac{P}{Q}}{\partial x}$ and for $n \in \mathbb{N}$, $f_Q^n(P)$ the same operation applied n times. We have then :*

$$f_Q^n(P) = \sum_{k=0}^{n} (-1)^k \sum_{i_0+\ldots+i_k=n-k} \frac{P^{(i_0)}(x)Q^{(i_1+1)}(x)\ldots Q^{(i_k+1)}(x)}{Q(x)^{n+k}} \sum_{\substack{0<n_1\leq i_0+1\\ \cdots\\ n_{k-1}<n_k<n}} \prod_{j=1}^{k} (n_j + j).$$

Proof. We will proceed by induction. It's to be noted that for $n > max(deg(Q), deg(P))$, many of the terms in the sum will be null, but we still denote them as a derivative of a certain order of Q or P

$$\mathcal{P}_n : \text{"} f_Q^n(P) = \sum_{k=0}^{n} (-1)^k \sum_{i_0+\ldots+i_k=n-k} \frac{P^{(i_0)}(x)Q^{(i_1+1)}(x)\ldots Q^{(i_k+1)}(x)}{Q(x)^{n+k}} \sum_{\substack{0<n_1\leq i_0+1\\ \cdots\\ n_{k-1}<n_k<n}} \prod_{j=1}^{k} (n_j + j)\text{"}.$$

If $n = 1$:

$$\begin{aligned} f(P) &= \frac{\partial \frac{P}{Q}}{\partial x} \\ &= \frac{P'}{Q} - \frac{PQ'}{Q^2} \\ &= (-1)^0 \frac{P^{(1)}}{Q} + (-1)^1 \frac{PQ^{(1)}}{Q^2}. \end{aligned}$$

We have the first step. Suppose we have $n \in \mathbb{N}$ such that $\mathcal{P}_n$ is true. Let's prove that $\mathcal{P}_{n+1}$ is also true.

$$\begin{aligned} f^{n+1}(P) =& f(f^n(P)) \\ =& \left(\sum_{k=0}^{n} (-1)^k \sum_{i_0+\ldots+i_k=n-k} \frac{P^{(i_0)}(x)Q^{(i_1+1)}(x)\ldots Q^{(i_k+1)}(x)}{Q(x)^{n+k+1}} \sum_{0<n_1\leq i_0+1,\ldots,n_{k-1}<n_k\leq k+\sum_{j=0}^{k} i_j} \prod_{j=1}^{k} (n_j + j) \right)' \end{aligned}$$

by hypothesis

$$= \sum_{k=0}^{n} (-1)^k \sum_{i_0+\ldots+i_k=n-k} \left(\frac{P^{(i_0)}(x)Q^{(i_1+1)}(x)\ldots Q^{(i_k+1)}(x)}{Q(x)^{n+k+1}} \right)' \sum_{0<n_1\leq i_0+1,\ldots,n_{k-1}<n_k\leq k+\sum_{j=0}^{k} i_j} \prod_{j=1}^{k} (n_j + j).$$

But we have :

$$\begin{aligned} \left(\frac{P^{(i_0)}(x)Q^{(i_1+1)}(x)\ldots Q^{(i_k+1)}(x)}{Q(x)^{n+k+1}} \right)' =& \frac{(P^{(i_0)}(x)Q^{(i_1+1)}(x)\ldots Q^{(i_k+1)}(x))'Q(x)^{(n+k+1)}}{Q(x)^{2(n+k+1)}} \\ &- \frac{(n+k+1)Q'Q^{n+k}(P^{(i_0)}(x)Q^{(i_1+1)}(x)\ldots Q^{(i_k+1)}(x))}{Q(x)^{2(n+k+1)}} \\ =& \frac{(P^{(i_0)}(x)Q^{(i_1+1)}(x)\ldots Q^{(i_k+1)}(x))'}{Q(x)^{n+k+1}} \\ &- \frac{(n+k+1)Q'(x)(P^{(i_0)}(x)Q^{(i_1+1)}(x)\ldots Q^{(i_k+1)}(x)}{Q(x)^{n+k+2}}). \end{aligned}$$

Moreover, focusing on $(P^{(i_0)}(x)Q^{(i_1+1)}(x)\ldots Q^{(i_k+1)}(x))'$:

$$(P^{(i_0)}(x)Q^{(i_1+1)}(x)\ldots Q^{(i_k+1)}(x))' = \sum_{j=0}^{k} P^{(i_0)}(x)Q^{(i_1+1)}(x)\ldots Q^{(i_j+2)}\ldots Q^{(i_k+1)}(x).$$

So we have :

$$\left(\frac{P^{(i_0)}(x)Q^{(i_1+1)}(x)\dots Q^{(i_k+1)}(x)}{Q(x)^{n+k+1}}\right)' = \sum_{j=0}^{k} \frac{P^{(i_0)}(x)Q^{(i_1+1)}(x)\dots Q^{(i_j+2)}\dots Q^{(i_k+1)}(x)}{Q(x)^{n+k+1}} - \frac{(n+k+1)Q'(x)(P^{(i_0)}(x)Q^{(i_1+1)}(x)\dots Q^{(i_k+1)}(x)}{Q(x)^{n+k+2}}).$$

Which with the entire sum gives :

$$\begin{aligned}
&\sum_{i_0+\dots+i_k=n-k} \sum_{j=0}^{k} \left(\frac{P^{(i_0)}(x)Q^{(i_1+1)}(x)\dots Q^{(i_j+2)}\dots Q^{(i_k+1)}(x)}{Q(x)^{n+k+1}}\right) \\
&- \frac{(n+k+1)Q'(x)(P^{(i_0)}(x)Q^{(i_1+1)}(x)\dots Q^{(i_k+1)}(x)}{Q(x)^{n+k+2}} \\
=& \sum_{i_0+\dots+i_k=n+1-k} \frac{P^{(i_0)}(x)Q^{(i_1+1)}(x)\dots Q^{(i_k+1)}(x)}{Q(x)^{n+k+1}} \\
&- \sum_{i_0+\dots+i_k=n-k} \frac{(n+k+1)Q'(x)(P^{(i_0)}(x)Q^{(i_1+1)}(x)\dots Q^{(i_k+1)}(x)}{Q(x)^{n+k+2}}.
\end{aligned}$$

Plugging the second sum into the whole expression we get :

$$\sum_{k=0}^{n}(-1)^{k+1} \sum_{i_0+\dots+i_k=n-k} \frac{Q'(x)(P^{(i_0)}(x)Q^{(i_1+1)}(x)\dots Q^{(i_k+1)}(x))}{Q(x)^{n+k+2}} \sum_{\substack{0<n_1\leq i_0+1 \\ \dots \\ n_{k-1}<n_k<n}} (n+k+1)\prod_{j=1}^{k}(n_j+j).$$

Taking $k' = k+1$:

$$= \sum_{k'=1}^{n+1}(-1)^{k'} \sum_{i_0+\dots+i_{k'-1}=n+1-k'} \frac{Q'(x)(P^{(i_0)}(x)Q^{(i_1+1)}(x)\dots Q^{(i_{k'-1}+1)}(x))}{Q(x)^{n+k'+1}} \sum_{\substack{0<n_1\leq i_0+1 \\ \dots \\ n_{k'-2}<n_{k'-1}<n}} (n+k')\prod_{j=1}^{k'-1}(n_j+j).$$

The first sum being :

$$\sum_{k=0}^{n}(-1)^{k} \sum_{i_0+\dots+i_k=n+1-k} \frac{P^{(i_0)}(x)Q^{(i_1+1)}(x)\dots Q^{(i_k+1)}(x)}{Q(x)^{n+k+1}} \sum_{\substack{0<n_1\leq i_0+1 \\ \dots \\ n_{k-1}<n_k<n}} \prod_{j=1}^{k}(n_j+j).$$

The second one exactly complete it to top $n+1$. Indeed, the last terms of the second sum is the terms needed to complete at the rank $n+1$ the formula. Moreover, each term of the second sum complete the first sum for the case when $n_k = n$. So we have :

$$\sum_{k=0}^{n+1}(-1)^{k} \sum_{i_0+\dots+i_k=n+1-k} \frac{P^{(i_0)}(x)Q^{(i_1+1)}(x)\dots Q^{(i_k+1)}(x)}{Q(x)^{n+k+1}} \sum_{\substack{0<n_1\leq i_0+1 \\ \dots \\ n_{k-1}<n_k<n+1}} \prod_{j=1}^{k}(n_j+j).$$

□

We need, before going for the second big lemma, to calculate the derivative to any order of Q :

Lemma 5.7. *If we denote $Q(x) := -g\left(\frac{m-g}{\eta}\right)^{b-1}$ with $b > 2$ and $Q^{(n)}(x)$ the derivative of Q to the n-th order we have :*

$$Q^{(n)}(x) = \prod_{i=1}^{n-1}(b-i)_+(-1)^n \frac{(m-g)^{(b-n-1)_+}}{\eta^{b-1}}(nm-bg)^{\mathbb{1}\{b\neq n\}} b^{\mathbb{1}\{b=n\}}.$$

Proof. Let's proceed by induction. The induction hypothesis is that for all $n \in [\![2\,;b]\!]$ the following property is true :

$$P_n : \text{"}Q^{(n)}(x) = (-1)^n \frac{(m-x)^{(b-n-1)_+}}{\eta^{b-1}}(nm-bx)^{\mathbb{1}\{b\neq n\}} b^{\mathbb{1}\{b=n\}} \prod_{i=1}^{n-1}(b-i)_+\text{"}.$$

We have :

$$Q'(x) = \frac{(m-x)^{(b-2)}}{\eta^{b-1}}(bx-m).$$

For $n = 2$, $n < b$:

$$\begin{aligned}\frac{\partial Q'}{\partial x}(x) &= b\frac{(m-g)^{b-2}}{\eta^{b-1}} - (b-2)\frac{(m-g)^{b-3}}{\eta^{b-1}}\\ &= \frac{(m-g)^{b-3}}{\eta^{b-1}}(2bm+bg-b^2g-2m)\\ &= \frac{(m-g)^{b-3}}{\eta^{b-1}}(2m-bg)(b-1).\end{aligned}$$

For $n = 2$, $n = b$:

$$\frac{\partial Q'}{\partial x}(x) = \frac{b}{\eta^{b-1}}.$$

So the property holds for $n = 2$.

Now suppose we have some $n \in \mathrm{N}$ such that the induction hypothesis holds. Then we have for $n+1$, $n+1 < b$:

$$Q^{(n+1)}(x) = \partial((-1)^n \frac{(m-x)^{b-n-1}}{\eta^{b-1}}(nm-bx)\prod_{i=1}^{n-1}(b-i))/\partial x,$$

by induction hypothesis

$$\begin{aligned}=&(-1)^{n+1}\frac{(m-x)^{b-n-2}}{\eta^{b-1}}(bm-bx+(b-n-1)(nm-bg))\prod_{i=1}^{n-1}(b-i)\\ =&(-1)^{n+1}\frac{(m-x)^{b-n-2}}{\eta^{b-1}}((b-n-1)nm-(b-n)bx+bm)\prod_{i=1}^{n-1}(b-i)\\ =&(-1)^{n+1}\frac{(m-x)^{b-n-2}}{\eta^{b-1}}([bn-n^2-n+b]m-(b-n)bx)\prod_{i=1}^{n-1}(b-i)\end{aligned}$$

with $(n+1)(b-n) = bn - n^2 + b - n$

$$=(-1)^{n+1}\frac{(m-x)^{b-n-2}}{\eta^{b-1}}((n+1)m-bg)\prod_{i=1}^{n}(b-i).$$

If $n+1=b$:

$$Q^{(n+1)}(x) = \partial((-1)^n \frac{(nm-bx)}{\eta^{b-1}} \prod_{i=1}^{n-1} (b-i)_+)/\partial x$$

by induction hypothesis

$$=(-1)^{n+1} \frac{\prod_{i=0}^{n-1}(b-i)}{\eta^{b-1}}$$

as $b=n+1, b-n=1$:

$$=(-1)^{n+1} \frac{\prod_{i=0}^{n}(b-i)}{\eta^{b-1}}.$$

The part $n>b$ is trivial. □

Lemma 5.8. *For $l,k \in \mathbb{N}$, for any $N \in \mathbb{N}$, if we denote $Q(g) := -g\left(\frac{m-g}{\eta}\right)^{b-1}$ and $P_l(g) := \frac{g^l}{(m-g)^s}$, we have :*

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

Proof. First we have to cut it in half :

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

I_2 can be roughly bounded as it will be negligible :

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

For the first one, we note $Q(x) = -g\left(\frac{m-g}{\eta}\right)^{b-1}$. The derivative of Q is different from 0 on the whole interval $[0, \frac{m}{b+1}]$ as it is equal to $Q'(x) = \frac{(m-g)^{b-2}}{\eta^{b-1}}(bg-m)$. If we denote $P_l(g) = \frac{g^l}{(m-g)^s}$. With integration by part we recognize the pattern we studied earlier :

$$\begin{aligned}
&\int_0^{m/(b+1)} g^l \exp(-g\left(\frac{m-g}{\eta}\right)^{b-1})dg \\
=&\left[\frac{P_l(g)}{Q'(g)} \exp(Q(g))\right]_0^{\frac{m}{b+1}} - \int_0^{m/(b+1)} \frac{\partial \frac{P}{Q'}}{\partial x}(g) \exp(Q(g))dg \\
=&\left[\frac{P_l(g)}{Q'(g)} \exp(Q(g))\right]_0^{\frac{m}{b+1}} - \int_0^{m/(b+1)} f_{Q'}(P_l)(g) \exp(Q(g))dg,
\end{aligned}$$

iterating $N+1$ times we get (with f^0 as the identity):

$$=\underbrace{\sum_{n=0}^{N} \left[\frac{f_{Q'}^n(P_l(g))}{Q'(g)} \exp(Q(g))\right]_0^{\frac{m}{b+1}}}_{S} + (-1)^{N+1} \int_0^{m/(b+1)} f_{Q'}^{N+1}(P_l)(g) \exp(Q(g))dg.$$

We need to show that $\int_0^{m/(b+1)} f_{Q'}^{N+1}(P_l)(g)\exp(Q(g))dg$ is negligible in front of S. Using Lemma 5.6, we know that for any $n \in \mathbb{N}$:

$$f_{Q'}^n(P_l)(x)\exp(Q(x)) = \sum_{k=0}^{n}(-1)^k \sum_{\substack{i_0+\ldots+i_k=n-k\\ i_0\leq l\\ i_1\leq b-2\\ \vdots\\ i_k\leq b-2}} \frac{P^{(i_0)}(x)Q^{(i_1+2)}(x)\ldots Q^{(i_k+2)}(x)}{Q'(x)^{n+k}} \sum_{\substack{0<n_1\leq i_0+1\\ \ldots\\ n_{k-1}<n_k<n}} \prod_{j=1}^{k}(n_j+j).$$

Using Lemma 5.7, we have for any $i < b$

$$Q^{(i+2)}(x) = \prod_{s=1}^{i-1}(b-s)_+(-1)^{i+2}\frac{(m-x)^{b-i-3}}{\eta^{b-1}}((i+2)m-bg).$$

And using Leibniz rule we get for P:

$$P^{(i_0)} = \sum_{k=0}^{i_0}\binom{i_0}{k}\prod_{j=0}^{k}(l-j)\prod_{i=0}^{n-k}(s-j)g^{l-k}(m-g)^{-s-i_0+k}.$$

We can already bound the derivative of P^{i_0} by the fact that the integration is within $[0, m/(b+1)]$ range :

$$P^{(i_0)} \leq \sum_{k=0}^{i_0}\binom{i_0}{k}\prod_{j=0}^{k}(l-j)\prod_{i=0}^{i_0-k}(s-j)m^{l-s-i_0}(b+1)^{s+i_0-l}.$$

Then for a given set of $\mathcal{I} = \{i_0, i_1, ..., i_k\}$ with $\forall j \in [0, k], i_j > 0$ and $i_0 + i_1 + ... + i_k = n - k$:

$$\begin{aligned}
&\frac{P^{(i_0)}(x)Q^{(i_1+2)}(x)\ldots Q^{(i_k+2)}(x)}{Q'(x)^{n+k}}\\
&\leq \sum_{q=0}^{i_0}\binom{i_0}{q}\frac{(b+1)^{s+i_0-l}l!s!(b-1)!^k(m-x)^{kb+i_0-n-2k}\prod_{i\in\mathcal{I}}((i+2)m-bx)}{(l-q)!(s-q)!\eta^{k(b-1)}(b-i_1-2)!\ldots(b-i_k-2)!}\times\\
&\frac{m^{l-s-i_0}\eta^{n+k(b-1)}}{(m-x)^{(n+k)(b-2)}(m-bx)^{n+k}},
\end{aligned}$$

using the fact that as $x \in [0, m/b+1]$ we have $\frac{(i+2)m-bx}{m-bx} \leq (b+1)(i+1)+1 \leq (b+1)(n+1)+1$ and simplifying :

$$\begin{aligned}
&\leq \sum_{q=0}^{i_0}\binom{i_0}{q}\frac{(b+1)^{s+i_0-l}l!s!(b-1)!^k((b+1)(n+1)+1)^k}{(l-q)!(s-q)!(b-i_1-2)!\ldots(b-i_k-2)!}\times\frac{m^{l-s-i_0}\eta^n}{(m-x)^{n(b-1)-i_0}(m-bx)^n}\\
&\leq \sum_{q=0}^{i_0}\binom{i_0}{q}\frac{(b+1)^{s+i_0-l}l!(b-1)!^k((b+1)(n+1)+1)^k}{(l-i_0)!(b-i_1-2)!\ldots(b-i_k-2)!}\times\frac{\eta^n(b+1)^{nb-i_0}}{m^{nb-l+s}b^{n(b-1)-i_0}}.
\end{aligned}$$

It's to be noted that we can extend with no difficulty the bound to the case where one or several of the i_j are superior or equal to b : if it's strictly superior the bound is trivial as the quantity is 0, and if it's b we simply have a constant. This allow us to extend the result to the case where $b = 2$.

Using the fact that the preceding majoration is for any combination $\{i_0, i_1, ..., i_k\}$ with $\forall j \in [0, k], i_j > 0$ and $i_0 + i_1 + ... + i_k = N + 1 - k$, we have for a C sufficiently big that does not depend on m:

$$\begin{aligned} m^{(N+1)b-l+s-1}\left|\int_0^{m/(b+1)} f_{Q'}^{N+1}(P_l)(g)\exp(Q(g))dg\right| &\le C\frac{m^{(N+1)b-l+s-1}}{m^{(N+1)b-l+s}}\times\left|\int_0^{m/(b+1)}\exp(Q(g))dg\right| \\ &\le C\frac{1}{m}\times\left|\int_0^{m/(b+1)}\exp(-g)dg\right| \\ &\le C\frac{1}{m}\times(1-\exp(-m/(b+1))) \\ &\underset{m\to\infty}{\to} 0. \end{aligned}$$

This concludes the proof.

□

The following lemmas are the building blocks necessary to calculate all the quantities we are interested in after, and are consequences of Lemma 5.8:

Lemma 5.9.

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

Proof. Use Lemma 5.8 with $l=0, s=0$ and $n=3$. □

Lemma 5.10.

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

Proof. Use Lemma 5.8 with $l=1, s=0$ and $n=3$. □

Lemma 5.11.

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

Proof. Use Lemma 5.8 with $l=2, s=0$ and $n=3$. □

Lemma 5.12.

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

Proof. Use Lemma 5.8 with $l=0, s=b-1$ and $n=3$. □

Lemma 5.13.

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

Proof. Use Lemma 5.8 with $l=0, s=b-2$ and $n=3$. □

Lemma 5.14.

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

Proof. Use Lemma 5.8 with $l=0, s=2b-2$ and $n=3$. ☐

Lemma 5.15. *If $G\sim\mathcal{E}(1)$ and the conditionnal 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 we have :*

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

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

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

The survival function of G gives :

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

A_1 gives :

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

So we have :

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

Applying Lemma 5.9 then yield the results

□

5.6. Proof of lemma 3.3 :

Lemma 3.3.

$$\mathbb{E}\left[G|M>m\right] \underset{m\to\infty}{=} \frac{\eta^{b-1}}{m^{b-1}} + \frac{4\eta^{2b-2}(b-1)}{m^{1-2b}} + \frac{\eta^{3b-3}(b-1)(31b-22)}{m^{3b-1}} - \frac{2(b-1)(3b-2)(27b-23)\eta^{4b-4}}{m^{4b-1}} + o\left(\frac{1}{m^{4b-1}}\right),$$

Proof. Denoting by $\alpha := \mathbb{P}\left(G+\xi>m\right)$:

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

Focusing first on A_3 :

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

For A_2, we can apply lemma 5.10.

We still need an equivalent to normalize :

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

using the classic development for geometric series :

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

As we will use it several time in the future, we denote it by Δ. Multiplying the result of the Lemma 5.10 and this yields the results. □

5.7. Proof of lemma 3.2 :

Lemma 3.2.

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

Proof.

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

denoting $M(x) := max(0, m-x)$

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

Calculating first A_4 with an integration by part :

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

Taking it into the full integral :

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

The result of Lemma 5.10 gives an equivalent for A_5, Lemma 5.12 gives an equivalent for A_6. Summing yields the expected value not normalised by the probability:

$$\alpha\mathbb{E}\left[\xi|G+\xi>m\right] \underset{m\to\infty}{=} \frac{\eta^{2b-2}}{m^{2b-2}} + \frac{6\eta^{3b-3}(b-1)}{m^{3b-2}} + \frac{12\eta^{4b-4}(b-1)(4b-3)}{m^{4b-2}} + o\left(\frac{1}{m^{4b-2}}\right).$$

Mutliplying by Δ yields the result. $\square$