Internet Appendix for Financial Information Theory

Gabriel E. Cabrera Guzmán

Supplementary Results

This Internet Appendix collects three items that the main text gestures at but does not carry out. Appendix A gives the proofs of the five inequalities gathered in the main-text cheat sheet. Appendix B derives the Kelly–Kullback–Leibler wealth-destruction identity from first principles for arbitrary odds. Appendix C sketches the two practical estimators of mutual information used in the empirical literature — the finite-alphabet plug-in and the Kraskov–Stögbauer–Grassberger k-nearest-neighbor estimator — with their bias–variance trade-offs.

Appendix A. Proofs

The main text uses five inequalities repeatedly and states them without proof. This appendix fills in the standard arguments. Every proof reduces to the parent statement, Jensen's inequality (Section A.1).

A.1 Jensen's inequality

Let X be an integrable random variable on a convex domain and ϕ a concave function on that domain. Then

$$\mathbb{E}[\phi(X)] \;\leq\; \phi(\mathbb{E}[X]), \tag{A.1}$$

with equality if and only if X is almost surely constant or ϕ is affine on the support of X.

Proof. Concavity of ϕ implies the existence of a supporting hyperplane at every interior point of the domain. Choose the supporting hyperplane at $\mu := \mathbb{E}[X]$, whose normal is any element γ of the superdifferential $\partial\phi(\mu)$:

$$\phi(x) \;\leq\; \phi(\mu) + \gamma'(x - \mu) \quad \text{for all } x \text{ in the domain of } \phi.$$

Take expectations. The right-hand side becomes $\phi(\mu) + \gamma'(\mathbb{E}[X] - \mu) = \phi(\mu)$, and the inequality of Eq. (A.1) follows. Equality across the integration requires the supporting hyperplane to touch ϕ almost everywhere on the support of X, which forces either $X = \mu$ a.s. or ϕ to be affine on the support. □

A.2 Gibbs' inequality

For any two probability measures P and Q on the same measurable space with $P \ll Q$,

$$D(P \,\|\, Q) \;\geq\; 0, \tag{A.2}$$

with equality if and only if $P = Q$ almost everywhere.

Proof. Apply Jensen (Eq. (A.1)) to the concave function $\phi = \ln$ and the random variable $Y = q(X)/p(X)$ under P:

$$-D(P \,\|\, Q) \;=\; \mathbb{E}_P\!\left[\ln \tfrac{q(X)}{p(X)}\right] \;\leq\; \ln \mathbb{E}_P\!\left[\tfrac{q(X)}{p(X)}\right] \;=\; \ln \int q(x)\,dx \;=\; \ln 1 \;=\; 0.$$

Multiplying by -1 reverses the inequality and gives Eq. (A.2). Equality demands $q(X)/p(X)$ to be P-almost surely constant. Since it must integrate to one under P, that constant is one and $P = Q$ a.e. □

A.3 Log-sum inequality

For non-negative reals $a_1, \ldots, a_m$ and $b_1, \ldots, b_m$ with $\sum_i b_i > 0$,

$$\sum_{i=1}^{m} a_i \ln \tfrac{a_i}{b_i} \;\geq\; \Big(\sum_{i=1}^{m} a_i\Big) \ln \tfrac{\sum_i a_i}{\sum_i b_i}, \tag{A.3}$$

with the convention $0 \ln 0 = 0$ and $a \ln(a/0) = \infty$ if $a > 0$. Equality holds if and only if a_i/b_i is constant across i with $b_i > 0$.

Proof. Set $A = \sum_i a_i$ and $B = \sum_i b_i$. The inequality is trivial if $A = 0$. Otherwise define the probability vectors $p_i = a_i/A$ and $q_i = b_i/B$. Substitute $a_i = Ap_i$ and $b_i = Bq_i$ into the left-hand side of Eq. (A.3):

$$\sum_i Ap_i \ln \frac{Ap_i}{Bq_i} \;=\; A \sum_i p_i \ln \frac{p_i}{q_i} + A \ln \frac{A}{B} \;=\; A \cdot D(p \parallel q) + A \ln \frac{A}{B}.$$

Gibbs (Eq. (A.2)) makes $D(p \parallel q) \geq 0$, so the expression is at least $A \ln(A/B)$, which is the right-hand side of Eq. (A.3). Equality requires $p = q$ across all i with $b_i > 0$, that is, $a_i/b_i = A/B$ constant. □

A.4 Data-processing inequality

If random variables X, Y, Z form a Markov chain $X \to Y \to Z$ — that is, Z is conditionally independent of X given Y — then

$$I(X; Z) \;\leq\; I(X; Y). \tag{A.4}$$

Proof. Expand the mutual information $I(X; (Y, Z))$ two ways using the chain rule:

$$I(X; (Y, Z)) \;=\; I(X; Y) + I(X; Z \mid Y) \;=\; I(X; Z) + I(X; Y \mid Z).$$

The Markov assumption gives $I(X; Z \mid Y) = 0$, since given Y the pair (X, Z) is independent. Rearranging,

$$I(X; Y) \;=\; I(X; Z) + I(X; Y \mid Z) \;\geq\; I(X; Z),$$

because $I(X; Y \mid Z) \geq 0$ (Gibbs applied to each conditional). □

The data-processing inequality captures the intuition that no deterministic or stochastic transformation of Y can recover information about X that was lost in the passage from X to Y. In the information bottleneck construction of the main text, this is exactly why the

compressed representation T cannot be more predictive of Y than X itself.

A.5 Fano's inequality

Let X be a discrete random variable on alphabet $\mathcal{X}$ and Y be an observation such that a decoder $\widehat{X} = g(Y)$ predicts X with error probability $P_e = \Pr(\widehat{X} \neq X)$. Then

$$H(P_e) + P_e \ln(|\mathcal{X}| - 1) \;\geq\; H(X \mid Y), \tag{A.5}$$

where $H(P_e) = -P_e \ln P_e - (1 - P_e) \ln(1 - P_e)$ is the binary entropy.

Proof. Introduce the error indicator $E = \mathbf{1}\{\widehat{X} \neq X\}$. Since E is a deterministic function of the pair (X, Y), one has $H(E \mid X, Y) = 0$. Apply the chain rule to $H(E, X \mid Y)$ two ways:

$$H(X \mid Y) \;=\; H(E, X \mid Y) \;=\; H(E \mid Y) + H(X \mid E, Y).$$

Bound each term. The first is at most the marginal binary entropy: $H(E \mid Y) \leq H(E) = H(P_e)$. The second decomposes by conditioning on E:

$$H(X \mid E, Y) \;=\; (1 - P_e)\, H(X \mid Y, E = 0) + P_e\, H(X \mid Y, E = 1).$$

When $E = 0$, $X = g(Y)$ is a deterministic function of Y, so $H(X \mid Y, E = 0) = 0$. When $E = 1$, X takes one of the $|\mathcal{X}| - 1$ values distinct from $g(Y)$, hence $H(X \mid Y, E = 1) \leq \ln(|\mathcal{X}| - 1)$. Combining the two bounds gives Eq. (A.5). □

Fano's inequality is the converse to Shannon's channel-coding theorem and, by extension, to the Kelly–Shannon capacity identity of the main text. No decoder can drive its error probability below the level pinned down by the source entropy $H(X \mid Y)$, no matter how

clever the decoding function is.

Appendix B. The Kelly–KL Identity

The main text asserts, without proof, that a horse-race gambler who bets according to belief Q instead of the truth P suffers a per-period wealth-growth deficit of exactly $D(P \parallel Q)$ nats. This appendix derives the identity for arbitrary odds and characterizes the special role of pari-mutuel pricing.

B.1 Setup

A gambler faces a horse race with m outcomes labeled $X \in \mathcal{X} = \{1, \dots, m\}$. Outcomes are drawn i.i.d. from the true law P, with $P(X = x) = p(x)$. The bookmaker posts odds $o(x) \geq 0$, meaning a one-nat bet on outcome x returns $o(x)$ nats if x occurs and zero otherwise. The gambler divides current wealth into fractions $b(x) \geq 0$ with $\sum_x b(x) = 1$ and bets $b(x)$ of wealth on outcome x. After one period, wealth is multiplied by $o(X)b(X)$. After n i.i.d. periods,

$$S_n \;=\; S_0 \prod_{t=1}^{n} o(X_t)b(X_t), \tag{B.1}$$

and by the strong law of large numbers, the per-period log growth rate converges almost surely to

$$W(b) \;:=\; \mathbb{E}_P[\ln\{o(X)b(X)\}] \;=\; \mathbb{E}_P[\ln o(X)] + \sum_x p(x) \ln b(x). \tag{B.2}$$

The first term of Eq. (B.2) is a fixed quantity determined by the odds structure and does not depend on the strategy. Only the second term, $\sum_x p(x) \ln b(x)$, is under the gambler's control.

B.2 Optimal proportional betting

The growth rate Eq. (B.2) is maximized by *proportional betting* $b^\star(x) = p(x)$, regardless of the odds structure. The optimum equals

$$W^\star \;=\; \mathbb{E}_P[\ln o(X)] - H(P). \tag{B.3}$$

Proof. Only the second term of Eq. (B.2) depends on b. Maximize $\sum_x p(x) \ln b(x)$ over the simplex by Lagrangian methods: set $\mathcal{L}(b, \lambda) = \sum_x p(x) \ln b(x) - \lambda(\sum_x b(x) - 1)$. The stationarity condition $\partial\mathcal{L}/\partial b(x) = p(x)/b(x) - \lambda = 0$ gives $b(x) = p(x)/\lambda$. Summing over x pins $\lambda = 1$, so $b^\star(x) = p(x)$. Substituting back, $W^\star = \mathbb{E}_P[\ln o(X)] + \sum_x p(x) \ln p(x) = \mathbb{E}_P[\ln o(X)] - H(P)$. □

Two features are worth flagging. The optimal fractions $b^\star(x) = p(x)$ are independent of the odds $o(x)$ — they depend only on the true probabilities. Larger odds simply raise every strategy's expected log wealth by the same constant. The optimum growth rate does depend on odds: bettor wealth grows positively iff the *return odds* $\mathbb{E}_P[\ln o(X)]$ exceed the entropy $H(P)$.

B.3 Pari-mutuel odds as the fair benchmark

Pari-mutuel pricing sets $o(x) = 1/p(x)$ — odds are the reciprocal of the true probabilities.

Under pari-mutuel odds, the optimum growth rate is zero:

$$W^\star_{\text{pari-mutuel}} \;=\; \mathbb{E}_P[\ln(1/p(X))] - H(P) \;=\; H(P) - H(P) \;=\; 0.$$

Pari-mutuel odds are the fair-market baseline. The optimum gambler cannot

systematically grow wealth against nature when the bookmaker prices at the true probabilities. Sub-fair odds ($o(x) < 1/p(x)$ for at least one x) drive $W^\star < 0$; super-fair odds admit a positive growth rate that measures the size of the bookmaker's mispricing.

B.4 The identity

Let $W^\star = W^\star(p)$ denote the growth rate under proportional betting on the truth, and let $W(q) := W(b = q)$ denote the growth rate when the gambler bets according to belief q. Then

$$W^\star(p) - W(q) \;=\; D(P \,\|\, Q). \tag{B.4}$$

Proof. Substitute $b(x) = q(x)$ into Eq. (B.2) and subtract Eq. (B.3):

$$W^\star(p) - W(q) \;=\; (\mathbb{E}_P[\ln o(X)] - H(P)) - (\mathbb{E}_P[\ln o(X)] + \mathbb{E}_P[\ln q(X)]).$$

The odds terms cancel. Simplifying,

$$W^\star(p) - W(q) \;=\; -H(P) - \mathbb{E}_P[\ln q(X)] \;=\; \sum_x p(x) \ln p(x) - \sum_x p(x) \ln q(x) \;=\; D(P \,\|\, Q),$$

by the definition of the Kullback–Leibler divergence. □

Eq. (B.4) is remarkable for what does not appear in it. The odds structure $o(x)$ has vanished from the deficit, leaving only the divergence between the truth P and the gambler's belief Q. Whatever the bookmaker charges, betting proportional to a wrong belief burns wealth at a rate of $D(P \,\|\, Q)$ nats per period. This is the operational "distance" content of the Kullback–Leibler divergence used throughout Section 4 of the main text (Cover and Thomas 2006, Ch. 6).

The identity generalizes to the setting with side information Y by conditioning both terms on Y (Barron and Cover 1988):

$$W^{\star}(p \mid Y) - W(q \mid Y) \;=\; \mathbb{E}_Y[D(P_{X|Y} \,\|\, Q_{X|Y})],$$

which links wealth growth from side information to conditional divergence and, ultimately, to the mutual information $I(X;Y)$ that plays the role of channel capacity in the main text.

Appendix C. Estimating Mutual Information

Mutual information is a functional of an unknown joint distribution. Its use in empirical finance — non-linear dependence detection, transfer-entropy causality tests, information-bottleneck regularizers — requires an estimator. This appendix reviews the two workhorses.

C.1 Plug-in estimation on discrete alphabets

Given N i.i.d. samples $\{(X_i, Y_i)\}_{i=1}^{N}$ from a joint law on a finite alphabet $\mathcal{X} \times \mathcal{Y}$, form empirical frequencies

$$\hat{p}(x, y) \;=\; \tfrac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{X_i = x, Y_i = y\}, \tag{C.1}$$

and marginals $\hat{p}(x), \hat{p}(y)$ by summing. The *plug-in estimator* substitutes the empirical laws into the population formula:

$$\hat{I}_{\text{plug}}(X;Y) \;=\; \sum_{x,y} \hat{p}(x, y) \ln \frac{\hat{p}(x, y)}{\hat{p}(x)\hat{p}(y)}. \tag{C.2}$$

The plug-in estimator is consistent as $N \to \infty$ but carries a well-known upward bias of order $O(m_{XY}/N)$, where m_{XY} is the number of jointly observed (x, y) cells (Paninski 2003). Small samples on rich alphabets inflate the estimate: a dependence signal appears where none exists. Two practical mitigations are standard. The Miller–Madow correction subtracts $(m_{XY} - m_X - m_Y + 1)/(2N)$ from Eq. (C.2). Alternatively, the alphabet can be coarsened so that each cell has at least, say, five observations before frequencies are computed (Steuer et al. 2002).

C.2 The KSG estimator

For continuous X and Y, plug-in estimation is infeasible without density estimation. Kraskov et al. (2004) developed a k-nearest-neighbor estimator — KSG, after Kraskov, Stögbauer, and Grassberger — that avoids density modeling entirely. The idea builds on the Kozachenko–Leonenko entropy estimator (Kozachenko and Leonenko 1987), which uses the distance to the k-th neighbor as a local scale proxy for the density.

Fix a neighbor rank $k \geq 1$ and let $z_i = (x_i, y_i) \in \mathbb{R}^{d_X + d_Y}$. For each sample i,

- let ε_i be the max-norm distance from z_i to its k-th nearest neighbor in the joint sample;
- let $n_X(i)$ be the number of samples $j \neq i$ with $\|x_j - x_i\|_\infty < \varepsilon_i/2$;
- let $n_Y(i)$ be the number of samples $j \neq i$ with $\|y_j - y_i\|_\infty < \varepsilon_i/2$.

The KSG estimator is

$$\hat{I}_{\mathrm{KSG}}(X;Y) \;=\; \psi(k) + \psi(N) - \frac{1}{N}\sum_{i=1}^{N}\big[\psi(n_X(i)+1) + \psi(n_Y(i)+1)\big], \tag{C.3}$$

where ψ is the digamma function. Two properties make Eq. (C.3) attractive for empirical finance. First, it is *invariant* to strictly monotone transformations of the two marginals separately, so return-scale choices (level versus log return, raw versus rank) do not perturb the estimate. Second, it requires no binning or density model. The bias is $O(k/N)$ under weak regularity conditions (Gao et al. 2015), and the variance decreases in k until the local density approximation breaks down. Practical choices in the empirical literature use $k \in [3, 6]$ for clean data and $k \in [10, 20]$ for noisy regimes.

C.3 Practical guidance

Three points regularly separate a defensible estimate from a noisy one.

First, *standardize each marginal* to unit variance before running the estimator. KSG is rotation-invariant only within each marginal block, and equally scaled coordinates make the max-norm neighbor search behave as intended.

Second, *break ties*. Tied distances are common in financial returns rounded to the tick, and they distort the neighbor counts n_X, n_Y. Adding a tiny jitter to each observation (uniform noise of order 10^{-10} of the standard deviation) is the standard fix.

Third, *report confidence intervals via the stationary bootstrap* (Politis and Romano 1994). Financial time series are autocorrelated, so naive resampling underestimates the standard error of $\hat{I}$. A block-bootstrap with mean block length tuned by the autocorrelation range produces valid intervals for both Eq. (C.2) and Eq. (C.3).

For transfer entropy the same estimators apply after rewriting the definition given in the main text as a difference of two mutual informations:

$$T_{X\to Y} \;=\; I(Y_{t+1}\,;\,X_t^{(k)},\,Y_t^{(\ell)}) - I(Y_{t+1}\,;\,Y_t^{(\ell)}). \tag{C.4}$$

Each term is estimated by KSG on the appropriate embedded sample. The difference structure cancels part of the KSG bias, which makes Eq. (C.4) preferable to a direct conditional-MI estimator for transfer-entropy applications in noisy market data (Wibral et al. 2013).

References

Barron, A. R. and Cover, T. M. (1988). A Bound on the Financial Value of Information. *IEEE Transactions on Information Theory*, 34(5):1097–1100.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, 2 edition.

Gao, S., Ver Steeg, G., and Galstyan, A. (2015). Efficient Estimation of Mutual Information for Strongly Dependent Variables. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 277–286.

Kozachenko, L. F. and Leonenko, N. N. (1987). Sample Estimate of the Entropy of a Random Vector. *Problems of Information Transmission*, 23(2):95–101.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating Mutual Information. *Physical Review E*, 69(6):066138.

Paninski, L. (2003). Estimation of Entropy and Mutual Information. *Neural Computation*, 15(6):1191–1253.

Politis, D. N. and Romano, J. P. (1994). The Stationary Bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313.

Steuer, R., Kurths, J., Daub, C. O., Weise, J., and Selbig, J. (2002). The Mutual Information: Detecting and Evaluating Dependencies between Variables. *Bioinformatics*, 18(Suppl. 2):S231–S240.

Wibral, M., Pampu, N., Priesemann, V., Siebenhühner, F., Seiwert, H., Lindner, M., Lizier, J. T., and Vicente, R. (2013). Measuring Information-Transfer Delays. *PLoS ONE*, 8(2):e55809.