# The Distortion-Perception Tradeoff in Finite Channels with Arbitrary Distortion Measures


**Dror Freirich**
Technion - Israel
Institute of Technology
drorfrc@gmail.com

**Nir Weinberger**
Technion - Israel
Institute of Technology
nirwein@technion.ac.il

**Ron Meir**
Technion - Israel
Institute of Technology
rmeir@ee.technion.ac.il


## Abstract


Whenever inspected by humans, reconstructed signals should not be distinguished from real ones. Typically, such a high perceptual quality comes at the price of high reconstruction error. We study this distortion-perception (DP) tradeoff over finite-alphabet channels, for the Wasserstein-1 distance as the perception index, and an arbitrary distortion matrix. We show that computing the DP function and the optimal reconstructions is equivalent to solving a set of linear programming problems. We prove that the DP curve is a piecewise linear function of the perception index, and derive a closed-form expression for the case of binary sources.


## 1 Introduction

The reconstruction of a signal from degraded data is required in numerous settings across science and engineering. In systems whose outputs are inspected by human users, reconstructions should not be easily distinguished from signals typical to the source domain. Such a high *perceptual quality* is achieved when the distribution of restored signals is close to the real signal's distribution [11, 4]. However, low distance between these distributions generally comes at the price of poor reconstruction distortion and vice versa. This leads to a tradeoff between distortion and perception, first studied in [4] (for a detailed introduction to the distortion-perception tradeoff, see Appendix A.1). The central problem is thus to quantify the *distortion-perception (DP) function*, which is the minimal distortion possible for a certain level of perceptual quality. The DP problem was studied by various authors. Specifically, [7] studied the DP function in real spaces, for the MSE distortion and the Wasserstein-2 perception index. In discrete spaces, [12, Thm.7] characterized the special case of a binary source, for the Hamming distortion and the Total-Variation (TV) perception index.

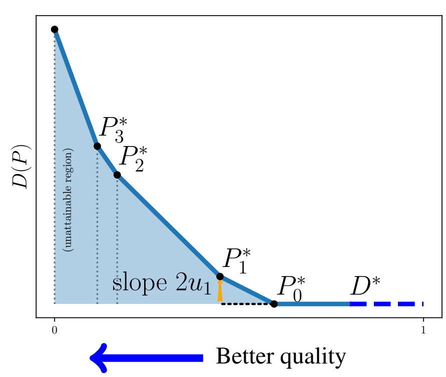


Figure 1: **The DP function.** $D(P)$ is a piecewise linear function. Breakpoints $P_i^*$ and slopes $2u_i$ are given explicitly by Theorem 4.2 for binary sources.

In this paper, we focus on discrete spaces, and investigate the DP tradeoff for general finite-alphabet channels and general distortion matrices. As the perception index, we consider the *Wasserstein*-1 distance induced by a general metric, which generalizes the TV distance [2, 16, 13]. We show that finding the DP function and the optimal reconstruction for this setting is equivalent to solving a set of *linear* problems, and the result is always a piecewise linear function of the perception index,


Information-Theoretic Principles in Cognitive Systems Workshop at the 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

regardless of the channel size, the underlying distributions or distortion measure. This stems from the properties of the *dual* feasible set. We further revisit the binary setting of [12], and derive a closed-form expression for the DP function, now considering a general distortion measure. We provide a self-contained proof for this case based on our novel analysis of the general setting.

## 2 Problem formulation

Let $X, Y$ be discrete variables defined on finite alphabets $\mathcal{X} = \{x_1, \dots, x_{n_x}\}, \mathcal{Y} = \{y_1, \dots, y_{n_y}\}$, where $X$ is the variable of interest, and $Y$ is a measurement of $X$ over a noisy channel. Their joint probability $p_{X,Y} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ is represented by the matrix $\mathbf{P}_{X,Y} = \{p(x, y)\}_{x,y \in \mathcal{X} \times \mathcal{Y}} \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{Y}|}$, and the marginal distributions $p_X$ and $p_Y$ are given by the vectors $\mathbf{P}_X \in \mathbb{R}^{|\mathcal{X}|}, \mathbf{P}_Y \in \mathbb{R}^{|\mathcal{Y}|}$. We assume that for each letter in the channel's output, $p_Y(y_i) > 0$ (*i.e.*, we ignore unused symbols). A randomized estimator $\hat{X} \in \mathcal{X}$ of $X$ from $Y$ is defined by a stochastic transition matrix $\mathbf{Q} = \mathbf{Q}_{\hat{X}|Y} \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{Y}|}$ whose entries are the probabilities $q(\hat{x}|y)$ to reconstruct the symbol $\hat{x} \in \mathcal{X}$ given that the channel output is $Y = y \in \mathcal{Y}$. We assume the Markov relation where $X, \hat{X}$ are independent given $Y$. The arbitrary *distortion* matrix is given by $\mathbf{D} = \{d(x, \hat{x})\}_{x,\hat{x} \in \mathcal{X}^2} \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{X}|}$, where the expected distortion $\mathbb{E}_{\mathbf{Q}}\left[d(X, \hat{X})\right] = \mathrm{Tr}\left\{\mathbf{P}_{X,Y}^\top \mathbf{D}\mathbf{Q}\right\}$ should be minimized w.r.t. $q(\hat{x}|y), \hat{x}, y \in \mathcal{X} \times \mathcal{Y}$. The marginal distribution $p_{\hat{X}}$ of $\hat{X}$ is given by the vector $\mathbf{P}_{\hat{X}} = \mathbf{Q}\mathbf{P}_Y$. We are interested in analyzing the *distortion-perception* (DP) *function* [4] $D(P) \triangleq \min_{\mathbf{Q}_{\hat{X}|Y}} \left\{\mathbb{E}_{\mathbf{Q}}\left[d(X, \hat{X})\right] \,:\, d_p(p_X, p_{\hat{X}}) \leq P\right\}$, where $d_p(\cdot, \cdot)$ denotes a distance between probability measures.

For simplicity, let us first consider the TV distance as the perceptual index $d_p$,

$$d_{TV}(\mathbf{P}_X, \mathbf{P}_{\hat{X}}) \triangleq \frac{1}{2} \sum_{x \in \mathcal{X}} |\mathbf{P}_X(x) - \mathbf{P}_{\hat{X}}(x)| = \sup_{A \subseteq \mathcal{X}} |p_X(A) - p_{\hat{X}}(A)|. \tag{1}$$

Note that $d_{TV}(\mathbf{P}_X, \mathbf{P}_{\hat{X}}) \in [0, 1]$, and $d_{TV}(\mathbf{P}_X, \mathbf{P}_{\hat{X}}) = 0$ iff $\mathbf{P}_X = \mathbf{P}_{\hat{X}}$. Then,

$$D(P) = \min_{\mathbf{Q}} \left\{ (\mathbf{D}^\top \mathbf{P}_{X,Y}) \bullet \mathbf{Q} : \mathbf{Q} \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{Y}|}, \begin{array}{c} \mathbf{1}^{|\mathcal{X}|} \cdot \mathbf{Q} = \mathbf{1}^{|\mathcal{Y}|} \\ d_{TV}(\mathbf{P}_X, \mathbf{Q}\mathbf{P}_Y) \leq P \end{array}, \mathbf{Q} \geq 0 \right\}, \tag{2}$$

where the Frobenius inner product $A \bullet B = \mathrm{Tr}\left\{A^\top B\right\}$, $\mathbf{1}^d$ is the $1 \times d$ dimensional all-ones vector, and $\mathbf{Q} \geq 0$ is applied elementwise. We start by presenting some elementary properties of (2).

**Proposition 2.1.** *Let $P \in [0, 1]$. The optimization problem* (2) *is feasible (namely, the constraints are satisfiable), and its optimal value is bounded from below.*

*Proof.* The *posterior sampling* solution $\mathbf{Q} = \mathbf{P}_{X|Y} = \{p_{X,Y}(x, y)/p_Y(y)\}_{x,y \in \mathcal{X} \times \mathcal{Y}}$ is feasible for every $P \geq 0$, since $\mathbf{P}_{\hat{X}} = \mathbf{Q}\mathbf{P}_Y = \mathbf{P}_X$, yielding $d_{TV}(\mathbf{P}_{\hat{X}}, \mathbf{P}_X) = 0$. For every stochastic matrix $\mathbf{Q}$, $(\mathbf{D}^\top \mathbf{P}_{X,Y}) \bullet \mathbf{Q} \in [\min D_{x,\hat{x}}, \max D_{x,\hat{x}}]$, hence the optimal value is bounded. □

**Proposition 2.2.** *Denote the matrix $\rho \triangleq \mathbf{D}^\top \mathbf{P}_{X,Y}$, whose entries are given by $\rho_{\hat{x},y} = \mathbf{P}_Y(y)\mathbb{E}\left[d(X, \hat{x})|Y = y\right]$. For any $P \geq 1$, $D(P) = \sum_y \min_{\hat{x} \in \mathcal{X}} \rho_{\hat{x},y} \triangleq D^*$. A corresponding optimal estimator is given by $\hat{X}^*(Y) \in \mathrm{argmin}_{\hat{x}}\, \rho_{\hat{x},Y}$. Trivially, $D(P) \geq D^*$ for every $P \in [0, 1]$.* The proof is straightforward.

## 3 Linear Programming formulation

We now observe that the perceptual constraint $\frac{1}{2} \sum_{x \in \mathcal{X}} |\mathbf{P}_X(x) - \sum_{y \in \mathcal{Y}} \mathbf{P}_Y(y)\mathbf{Q}(x|y)| \leq P$ in (2), is equivalent to the set of *linear* constraints

$$\sum_{x \in \mathcal{X}} \pm(\mathbf{P}_X(x) - \sum_{y \in \mathcal{Y}} \mathbf{P}_Y(y)\mathbf{Q}(x|y)) \leq 2P. \tag{3}$$

Taking all possible sign combinations we attain $2^{|\mathcal{X}|}$ linear constraints, where the 2 constraints for which the signs are either all positive or all negative are redundant since the LHS of (3) vanishes.

Together with (2), we can reformulate the DP function as the following Linear Program (LP) [3, 17]

$$D(P) = \min_{\mathbf{Q}\geq 0}\left\{\rho \bullet \mathbf{Q} : \begin{array}{c} \mathbf{1}^{|\mathcal{X}|}\cdot \mathbf{Q} = \mathbf{1}^{|\mathcal{Y}|},\ \mathbf{Q}\in\mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|} \\ \sum_{x\in\mathcal{X}} \pm\left(\mathbf{P}_X(x) - \sum_{y\in\mathcal{Y}} \mathbf{P}_Y(y)\mathbf{Q}_{x|y}\right) \leq 2P \end{array}\right\}. \quad (4)$$

In (4), we have $|\mathcal{X}|\times|\mathcal{Y}|$ variables (the entries of $\mathbf{Q} = \{q(\hat{x}|y)\}$), and $|\mathcal{Y}| + 2^{|\mathcal{X}|} - 2$ constraints.

**Total Variation as Optimal Transport** Let $\mathbf{H} = \{1 - \delta_{x,\hat{x}}\}_{x,\hat{x}\in\mathcal{X}\times\mathcal{X}}$ be the *Hamming* distance matrix, let $\mathscr{P}(\mathcal{X})$ be the set of probability measures on $\mathcal{X}$, and let $\Pi \in \mathscr{P}(\mathcal{X}\times\mathcal{X})$ have marginals $\mathbf{P}_X$ and $\mathbf{P}_{\hat{X}}$ (parameterized by a matrix $\mathbf{\Pi}_{\mathbf{x},\hat{\mathbf{x}}}$). It is well known [16] that taking $\mathbf{H}$ as a metric on $\mathcal{X}$, the TV distance coincides with the *Wasserstein*-1 distance on $\mathscr{P}(\mathcal{X})$, namely

$$d_{TV}(\mathbf{P}_X, \mathbf{P}_{\hat{X}}) = W_{1,H}(\mathbf{P}_X, \mathbf{P}_{\hat{X}}) \triangleq \inf_{\Pi} \Pi\left[x \neq \hat{x}\right] = \inf_{\mathbf{\Pi}} \mathbf{\Pi}\bullet\mathbf{H}, \quad (5)$$

where the minimum is attained [13, Lemma 3.4.1]. Wasserstein distances are convex metrics on $\mathcal{P}(\mathcal{X})$ [2]. Using (5), we can rewrite (4) as the linear problem

$$D(P) = \min_{\substack{\mathbf{Q},\mathbf{\Pi},\\ \varepsilon\geq 0}}\left\{\rho\bullet\mathbf{Q} : \begin{array}{cc} \sum_{\hat{x}\in\mathcal{X}} \mathbf{P}_Y(y)\mathbf{Q}_{\hat{x}|y} = \mathbf{P}_Y(y), \forall y\in\mathcal{Y} & \mathbf{Q}\in\mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|} \\ \sum_{\hat{x}\in\mathcal{X}} \mathbf{\Pi}_{x,\hat{x}} = \mathbf{P}_X(x), \forall x\in\mathcal{X} & \mathbf{\Pi}\in\mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|} \\ \sum_{x\in\mathcal{X}} \mathbf{\Pi}_{x,\hat{x}} = \sum_{y\in\mathcal{Y}} \mathbf{P}_Y(y)\mathbf{Q}_{\hat{x}|y}, \forall\hat{x}\in\mathcal{X} & \mathbf{\Pi}\bullet\mathbf{H} + \varepsilon = P \end{array}\right\}, \quad (6)$$

where $\varepsilon$ is a slack variable. The problem (6) possesses $|\mathcal{X}|(|\mathcal{Y}| + |\mathcal{X}|) + 1$ variables and only $|\mathcal{Y}| + 2|\mathcal{X}| + 1$ constraints, from which $|\mathcal{Y}| + 2|\mathcal{X}|$ are independent. Interestingly, the form (6) allows to discuss a more general family of perceptual divergences − Wasserstein-1 distances induced by arbitrary metrics $H$ on $\mathcal{X}$, which we will consider to be the case from this point on. We will assume *w.l.o.g.* that $H$ takes values in $[0, 1]$, hence the results of Propositions 2.1 and 2.2 hold trivially in this case.

**The Dual Problem** Let the general linear programming problem be [3]

$$\text{(LP)} \quad \min_q z^\top q, \text{ s.t. } \mathbf{A}q = b \text{ and } q \geq 0, \qquad q, z\in\mathbb{R}^n, b\in\mathbb{R}^{n_c}, \mathbf{A}\in\mathbb{R}^{n_c\times n}, \quad (7)$$

and the inequality is elementwise. Its dual problem (DLP) is given by

$$\text{(DLP)} \quad \max_w w^\top b, \text{ s.t. } w^\top\mathbf{A} \leq z^\top. \quad (8)$$

*Strong duality* holds for feasible and bounded LP problems [3], namely, the problem (8) is feasible and $\min_q z^\top q = \max_w w^\top b$. We next derive the dual form of (6). For convenience, we split the variables in (8) into four groups: $|\mathcal{Y}|$ variables $\{w_y\}_{y\in\mathcal{Y}}$ related to the stochasticity constraint on $\mathbf{Q}$ for each symbol in $\mathcal{Y}$, the two groups of $|\mathcal{X}|$ variables $\{r_x\}$ and $\{\nu_{\hat{x}}\}$ related to the constraints on the marginals of $\mathbf{\Pi}$, and the variable $l$ related to the perception constraint $\mathbf{\Pi}\bullet\mathbf{H} + \varepsilon = P$. We denote

$$\rho'_{\hat{x},y} \triangleq \frac{\rho_{\hat{x},y}}{\mathbf{P}_Y(y)} = \mathbb{E}\left[d(X,\hat{x})|Y = y\right], \quad (9)$$

and explicitly write the dual problem of (6) as (see derivation in the Appendix),

$$\max_{w,r,\nu,l}\left[\sum_{y\in\mathcal{Y}} p_y w_y + \sum_{x\in\mathcal{X}} p_x r_x - lP\right] \text{ s.t. } \left\{ l \geq 0, \begin{array}{ll} w_y \leq \rho'_{\hat{x},y} - \nu_{\hat{x}}, & \forall\hat{x}, y\in\mathcal{X}\times\mathcal{Y} \\ r_x \leq \mathbf{H}_{x,\hat{x}}l + \nu_{\hat{x}}, & \forall x,\hat{x}\in\mathcal{X}\times\mathcal{X} \end{array}\right. . \quad (10)$$

Given a value $P$, $D(P)$ can be calculated by numerically solving (6) (equivalently, (10)). However, finding a closed-form solution remains an open problem. In Section 4.2 we find such an expression for small problems. We further observe that the objective of (10) is linear in the perception index, hence the maximal value for a given $P$ is attained by some non-increasing linear function of the form $p_0 - p_1 P$ . We further develop this insight below.

# 4 Main results

## 4.1 Piecewise linearity of DP functions

While the problem of finding an exact formula for $D(P)$ is still open, here we exploit the properties of the dual problem (10) in order to show the general property that $D(P)$ is piecewise linear in the perception index $P$. Moreover, as we show in Appendix B, the breakpoints and slopes of this function are determined by the vertices of a convex set in $\mathbb{R}^2$.

**Theorem 4.1.** *For $P \in [0, \infty)$, the DP function* (6) *is a non-increasing, piecewise linear function of $P$, with a non-decreasing slope. Also, there exists $P^* \in [0, 1]$ such that $D(P) = D^*$ for all $P \geq P^*$.*

The proof is based on analyzing the dual formulation (10). Due to strong duality this matches the primal problem. The feasible set of (10) has a finite number of vertices, and this set is independent of the perceptual index $P$. The solution to (10) must occur at one of these vertices. Thus, the interval $[0, 1]$ may be partitioned into sub-intervals, so that in each sub-interval the solution to (10) is at the same vertex. For a fixed choice of variables $w, r, \nu$ and $l$ in (10), the $D(P)$ function is linear with slope $-l$. Hence, the DP function is piecewise linear. Since DP functions are non-increasing and convex (see Thm. A.1 in the Appendix), the slope cannot decrease. A full proof can be found in the Appendix for the case of the TV perceptual index. We note, however, that the form (10) implies that the same arguments hold where the perceptual distance is replaced by the Wasserstein-1 distance (5) induced by any metric $H$ on $\mathcal{X}$ (taking values in $[0, 1]$).

### 4.2 Full characterization of channels with binary sources

We next focus on the case of binary sources, where $\mathcal{X} = \{x_1, x_2\}$ with probabilities $p_{x_1}, p_{x_2}$, respectively, and $\mathcal{Y}$ is of arbitrary size $n_y$. It suffices to analyze the TV distance (1) as the perceptual index, since every metric defining the Wasserstein-1 distance is proportional to the Hamming distance in the binary case. The distortion matrix is arbitrary, yielding the matrix $\rho'$ defined in (9). Denote $u_y = \frac{1}{2}(\rho'_{\hat{x}_1 y} - \rho'_{\hat{x}_2 y})$ which is half the cost of reconstructing $y$ as $x_1$ over reconstructing as $x_2$, and assume *w.l.o.g.* that $u_{y_1} \leq u_{y_2} \leq \ldots \leq u_{y_n}$. We define $P_Y^-(u) = \Pr\{u_Y \leq u\} = \sum_{y: u_y \leq u} \mathbf{P}_Y(y)$, which is right-continuous with left limit $P_Y^-(u^-) = \Pr\{u_Y < u\} = \sum_{y: u_y < u} \mathbf{P}_Y(y)$. We further denote the symbols $y_i^*$ whose $u_y$ is non-zero, namely

$$u_{-M^-} = u_{y^*_{-M^-}} \leq \ldots \leq u_{-1} = u_{y^*_{-1}} < 0 = u_0 < u_1 = u_{y^*_1} \leq \ldots \leq u_{M^+} = u_{y^*_{M^+}}. \tag{11}$$

**Theorem 4.2.** *Assume that $p_{x_1} \geq P_Y^-(0)$, and let $I = \max\{i \colon p_{x_1} \geq P_Y^-(u_i)\}$. Then, the DP function $D(P)$ is piecewise linear with breakpoints $\{P_i^*\}_{i=0}^I$ given by*

$$P_i^* = p_{x_1} - P_Y^-(u_i) \tag{12}$$

*where, specifically, $P_0^* = p_{x_1} - P_Y^-(0) = P^*$. The DP function is then given by*

$$D(P) = \begin{cases} D^*, & P \geq P_0^* \\ D(P_{i-1}^*) + 2u_i \left(P_{i-1}^* - P\right), & P_i^* \leq P \leq P_{i-1}^* \\ D(P_I^*) + 2u_{I+1} \left(P_I^* - P\right), & 0 \leq P \leq P_I^* \end{cases}. \tag{13}$$

*If $P_Y^-(0^-) \geq p_{x_1}$, then similarly $P_0^* = P_Y^-(0^-) - p_{x_1}$, and $P_i^* = P_Y^-(u_{-i-1}) - p_{x_1}$, while it is non-negative, and $D(P)$ is determined analogously. In the case $P_Y^-(0) \geq p_{x_1} \geq P_Y^-(0^-)$, $P^* = 0$ and $D(P) \equiv D^*$ for all $P \geq 0$.*

*Remark* 4.3. If $u_i = u_{i-1}$ then $P_i^* = P_{i-1}^*$ and this yields a 'degenerate' interval. If $u_i > u_{i-1}$, then (12) can alternatively be written more simply as $P_i^* = P_{i-1}^* - \mathbf{P}_Y(y_i^*)$.

The results of Theorem 4.2 are illustrated in Fig. 1. These results reassure the intuition that channel outputs in $\mathcal{Y}$ should be mapped to symbols in $\{x_1, x_2\}$ in a *greedy* fashion; At the point $P = 1$, each $y$ is reconstructed with a minimal penalty, without any perceptual constraints (as in Proposition 2.2). This can be done by setting, *e.g.*, $q(\hat{x}_1|y) = \delta_{u_y \leq 0}$ . At the point $P = P^*$, $y$'s are still reconstructed optimally, but now under a perception constraint. This can be obtained by rearranging the mapping of symbols whose $u_y = 0$, which yields no extra cost in distortion. Now, suppose that $x_1$ is not 'fully allocated', that is, $p_{x_1} \geq P_Y^-(0)$. As the perception constraint becomes more restrictive (lower $P$), the estimator will seek for the minimal cost symbols $y \in \mathcal{Y}$ that are mapped to $x_1$ with probability less than 1, and increase this probability. For a small change of $\Delta P$, the cost in distortion is $2u_y \Delta P$. This is done until $P = 0$ is met, namely $p_{\hat{x}_1} = p_{x_1}$.

**Corollary 4.4.** *At the breakpoints where $P_i^* \neq 0$, an optimal estimator is given by a deterministic rule $Q_{P_i^*}$ (for $p_{x_1} \geq P_Y^-(0)$, given by $Q_{P_i^*} = \left\{q(x_1|y) = \delta_{u_y \leq u_i}\right\}$). Interestingly, at $P \in \left[P_i^*, P_{i-1}^*\right]$, the estimator is given by the convex combination of estimators at the interval edges, $Q_P = \alpha Q_{P_{i-1}^*} + (1 - \alpha) Q_{P_i^*}$, with $\alpha = \frac{P - P_i^*}{P_{i-1}^* - P_i^*}$. This result implies that in order to construct an estimator for any point along the tradeoff at test time, without any additional calculations, it is sufficient to calculate $\mathcal{O}(|\mathcal{Y}|)$ estimators beforehand, one at each breakpoint (and at $P = 0$).*

**Acknowledgements** This research was supported by the Israel Science Foundation (ISF), grant no. 1782/22. The work of RM was partially supported by the Skillman chair in biomedical sciences and by the Ollendorff Minerva Center, ECE Faculty, Technion.

## References

[1] Theo Adrai, Guy Ohayon, Tomer Michaeli, and Michael Elad. Deep optimal transport: A practical algorithm for photo-realistic image restoration. *arXiv preprint arXiv:2306.02342*, 2023.

[2] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[3] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

[4] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018.

[5] Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, pages 675–685. PMLR, 2019.

[6] Jun Chen, Lei Yu, Jia Wang, Wuxian Shi, Yiqun Ge, and Wen Tong. On the rate-distortion-perception function. *arXiv preprint arXiv:2204.06049*, 2022.

[7] Dror Freirich, Tomer Michaeli, and Ron Meir. A theory of the distortion-perception tradeoff in wasserstein space. *Advances in Neural Information Processing Systems*, 34:25661–25672, 2021.

[8] Dror Freirich, Tomer Michaeli, and Ron Meir. Perceptual kalman filters: Online state estimation under a perfect perceptual-quality constraint. *arXiv preprint arXiv:2306.02400*, 2023.

[9] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[10] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[11] Frank Nielsen. Hypothesis testing, information divergence and computational geometry. In *International Conference on Geometric Science of Information*, pages 241–248. Springer, 2013.

[12] Jingjing Qian, George Zhang, Jun Chen, and Ashish Khisti. A rate-distortion-perception theory for binary sources. In *International Zurich Seminar on Information and Communication (IZS 2022). Proceedings*, pages 34–38. ETH Zurich, 2022.

[13] Maxim Raginsky, Igal Sason, et al. Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends® in Communications and Information Theory*, 10(1-2):1–246, 2013.

[14] Sadaf Salehkalaibar, Buu Phan, Ashish Khisti, and Wei Yu. Rate-distortion-perception tradeoff based on the conditional perception measure. In *2023 Biennial Symposium on Communications (BSC)*, pages 31–37. IEEE, 2023.

[15] Lucas Theis and Aaron B Wagner. A coding theorem for the rate-distortion-perception function. *arXiv preprint arXiv:2104.13662*, 2021.

[16] Ramon Van Handel. Probability in high dimension. *Lecture Notes (Princeton University)*, 2014.

[17] Robert J Vanderbei et al. *Linear programming*. Springer, 2020.

[18] Aaron B Wagner. The rate-distortion-perception tradeoff: The role of common randomness. *arXiv preprint arXiv:2202.04147*, 2022.

[19] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[20] Zeyu Yan, Fei Wen, and Peilin Liu. Optimally controllable perceptual lossy compression. *arXiv preprint arXiv:2206.10082*, 2022.

# A Preliminaries

## A.1 The distortion-perception tradeoff

In this section we provide an overview on the distortion-perception (DP) tradeoff. The problem of reconstructing a signal from a corrupted measurement arises in many settings, such as decoding of transmitted communication signals and signal enhancement. Until recently, reconstruction algorithms' performance has been measured by its mean *distortion*, such as mean squared error (MSE). For that reason, many methods aimed to minnimize distortion measures such as MSE and peak signal-to-noise ratio (PSNR). However, in systems where outputs are inspected by human users, reconstructions should not be easily distinguished from signals typical to the source domain. Therefore, many current works target *perceptual quality* rather than distortion (*e.g.* in image restoration, see [1, 19, 10, 9]).

Mathematically, the probability of success in a hypothesis test is known to be proportional to the Total-Variation (TV) distance between distributions [11]. Hence, we may consider high perceptual quality to be achieved when the distribution of restored signals is close to the real signals distribution [4]. Good perceptual quality generally comes at the price of high reconstruction error and vice versa. This *perception-distortion tradeoff* was first studied in [4]. In the study of the distortion-perception (DP) tradeoff, a central problem is to quantify the the minimal distortion possible for a certain level of perceptual quality.

While in this paper we focus our efforts on discrete spaces, here we formulate the setting of general metric spaces for the sake of completeness.

Let $X, Y$ be random variables taking values in some complete separable metric spaces $\mathcal{X}$,$\mathcal{Y}$, respectively. We assume the existence of the joint probability $p_{X,Y}$ on $\mathcal{X} \times \mathcal{Y}$, and a Borel lower-bounded distortion function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+ \cup \{0\}$. An estimator $\hat{X} \in \mathcal{X}$ is a random variable on $\mathcal{X}$, defined by its distribution conditioned on the measurement $Y$, $p_{\hat{X}|Y}$, with marginal distribution $p_{\hat{X}}$.

An *optimal* estimator for the DP tradeoff, is an estimator that minimizes the expected *distortion* $\mathbb{E}[d(X, \hat{X})]$ under the *perception* constraint $d_p(p_X, p_{\hat{X}}) \leq P$. Here, $d_p$ is a divergence between probability measures. Blau and Michaeli [4] introduced the distortion-perception *function*

$$D(P) \triangleq \min_{p_{\hat{X}|Y}} \left\{ \mathbb{E}[d(X, \hat{X})] \;:\; d_p(p_X, p_{\hat{X}}) \leq P \right\}. \tag{14}$$

The expectation is taken *w.r.t* the probability measure induced by $p_{XY}$ and $p_{\hat{X}|Y}$ where we assume that $X, \hat{X}$ are independent given $Y$. We have the following result of Blau and Michaeli [4, Thm.2].

**Theorem A.1.** *(The perception-distortion tradeoff). If $d_p(p, q)$ is convex in its second argument (which is the case for the TV and Wasserstein distances discussed in this paper), then the distortion-perception function* (14) *is monotonically non-increasing and convex.*

Apart from the general properties above, the precise nature of DP functions depends on the exact setup. [7] fully characterises this function in real spaces, considering the MSE and Wasserstein-2 indices. Recently, [8] investigated the cost of perfect perceptual consistency in online estimation problems.

The DP tradeoff was extended to lossy compression by presenting the rate-distortion-perception (RDP) function [5, 14, 20, 12], which is the minimal rate of a code whose decoding allows a desired tradeoff between reconstruction and perceptual quality. A coding theorem was introduced for this setting [15, 18], where the properties of optimal codes are investigated [6].

## A.2 The linear optimization problem and strong duality

Let the general Linear Programming (LP) problem [3]

$$\begin{cases} \rho \bullet \mathbf{Q} & \to \min_{\mathbf{Q}} \\ \text{s.t.} & a_i \bullet \mathbf{Q} = b_i,\ i \in M_1 \ . \\ & s_i \bullet \mathbf{Q} \leq b_i,\ i \in M_2 \\ & \mathbf{Q} \geq 0 \end{cases} \tag{15}$$

$\mathbf{Q}, \rho, a_i$ are real $|\mathcal{X}| \times |\mathcal{Y}|$ matrices, $b = \{b_i\}_{i \in M} \in \mathbb{R}^{n_c}$. The Dual Linear Programming problem (DLP) is given by

$$\begin{cases} w^\top b & \to \max_w \\ \text{s.t.} & w_i \leq 0,\, i \in M_2 \\ & w_i \in \mathbb{R},\, i \in M_1 \\ & \sum_{i \in M_1} w_i \{a_i\}_{x,y} + \sum_{i \in M_2} w_i \{s_i\}_{x,y} \leq \rho_{x,y}, \\ & \forall x, y \in \mathcal{X} \times \mathcal{Y} \end{cases} . \tag{16}$$

Recall that by slight abuse of notation, here, similarly to the main text, we use $x = x_\alpha$ and $y = y_\beta$ to denote their indices $\alpha$ and $\beta$, respectively.

Dual problems are useful for establishing lower bounds on the optimal value, due to the property of *weak duality*, which assures that every feasible value for the Primal problem is greater than or equal to every feasible value of its Dual, yielding (in case where both problems are feasible)

$$\min_{\mathbf{Q}} \rho \bullet \mathbf{Q} \geq \max_w w^\top b. \tag{17}$$

For feasible, bounded LP problems we further possess a *strong duality*, namely the problem (16) is feasible and

$$\min_{\mathbf{Q}} \rho \bullet \mathbf{Q} = \max_w w^\top b. \tag{18}$$

#### A.2.1 A dual form for the TV distance setting

For our future analysis, here it is convenient to use the dual of the form (4) to $D(P)$. In this formulation, we have $\rho = \mathbf{D}^\top \mathbf{P}_{X,Y}$, and we can write (15)

$$b^\top = b(P)^\top = \left[p_{y_1}, \ldots, p_{y_n}, 2P - S_1^\top \mathbf{P}_X, \ldots, 2P - S_{2^{|\mathcal{X}|}-2}^\top \mathbf{P}_X\right], \tag{19}$$

where $P$ is the perception index. Also,

$$a_j = \mathbf{P}_Y(y_j) \mathbf{1}^{|\mathcal{X}|\top} e_j,\; j = 1, \ldots, |\mathcal{Y}|\,, \tag{20}$$

$$s_i = S_i \mathbf{P}_Y^\top,\; i = 1, \ldots, 2^{|\mathcal{X}|} - 2\,, \tag{21}$$

$S_i$ are the vectors of the set $\{-1, 1\}^{|\mathcal{X}|} \setminus \{\pm [1, ..., 1]\}$, and $e_j$ is the $j$-th unit vector in the standard basis.

For convenience, let us split the decision variables in (16) into two groups; $|\mathcal{Y}|$ variables $\{w_y\}_{y \in \mathcal{Y}}$ related to the stochasticity constraint for each symbol in $\mathcal{Y}$, and the $2^{|\mathcal{X}|} - 2$ variables $\{\nu_i\}$ related to the perception constraints (3). Now, (16) becomes

$$\begin{cases} [w^\top, \nu^\top] b & \to \max_{w,\nu} \\ \text{s.t.} & \nu_i \leq 0,\;\, i = 1, \ldots, 2^{|\mathcal{X}|} - 2 \\ & w_y \in \mathbb{R},\, y \in \mathcal{Y} \\ & \sum_j w_j \{a_j\}_{x,y} - \sum_i \nu_i \{s_i\}_{x,y} \leq \rho_{x,y},\, \forall x, y \in \mathcal{X} \times \mathcal{Y}, \end{cases} , \tag{22}$$

where $\{a_j\}$ is a matrix containing $\mathbf{P}_Y(y_j)$ along its $j$-th column, and 0 elsewhere. Hence, for every $y \in \mathcal{Y}$, only the corresponding $\{a_j\}$ (where $y_j = y$) contributes to the sum, namely $\sum_j w_j \{a_j\}_{x,y} = \mathbf{P}_Y(y) w_y$ regardless of $x$. In addition, $s_i$ is a matrix where the $j$-th column is given by $\mathbf{P}_Y(y_j) S_i$, thus $\{s_i\}_{x,y} = \mathbf{P}_Y(y) \{S_i\}_x$. Recall that we denote

$$\rho'_{\hat{x},y} = \frac{\rho_{\hat{x},y}}{p_Y(y)}, \tag{23}$$

and directly write (22) as

$$\begin{cases} \sum_y \mathbf{P}_Y(y) w_y & +2P \left(\sum_i \nu_i\right) - \left(\sum_i \nu_i S_i^\top\right) \mathbf{P}_X \to \max_{w,\nu} \\ \text{s.t.} & \nu_i \leq 0,\;\, i = 1, \ldots, 2^{|\mathcal{X}|} - 2 \\ & w_y \leq \rho'_{x,y} + \sum_i \nu_i \{S_i\}_x,\, \forall x, y \in \mathcal{X} \times \mathcal{Y} \end{cases} . \tag{24}$$

Note again that $\{S_i\}_x$ is the sign of $\mathbf{P}_X(x) - \sum_y \mathbf{Q}(x|y)\mathbf{P}_Y(y)$ in the $i$-th constraint of (3), and that $\sum_i \nu_i S_i^\top \mathbf{P}_x = \sum_x \mathbf{P}_X(x) \sum_i \nu_i \{S_i\}_x$. We can write the dual feasible set as

$$\mathcal{S} = \left\{ p_i = [w^i, \nu^i] \in \mathbb{R}^{|\mathcal{Y}|+R} : p_i^\top \mathbf{A} \le c \right\}, \tag{25}$$

where $R = 2^{|\mathcal{X}|} - 2$, $\otimes$ is the Kronecker product, and

$$\mathbf{A} = \left[ \begin{array}{c|c} 1^{|\mathcal{X}|} \otimes \mathrm{diag}\{\mathbf{P}_Y\} & 0_{|\mathcal{Y}| \times R} \\ \hline -S \otimes \mathbf{P}_Y^\top & I_R \end{array} \right] \in \mathbb{R}^{(|\mathcal{Y}|+R) \times (|\mathcal{X}||\mathcal{Y}|+R)}, \tag{26}$$

and

$$c = \left[ \mathrm{rowstack}\{\rho\}^\top, 0_{1\times R} \right] \in \mathbb{R}^{|\mathcal{X}||\mathcal{Y}|+R}. \tag{27}$$

Now, $S$ is a matrix whose $j$-th row is $S_j^T$. Since the Primal problem possesses a finite optimal solution for $P \in [0, \infty)$, its dual must be feasible and bounded (Bertsimas and Tsitsiklis [3, pp.151]). It is then easy to see that $\mathrm{rank}(\mathbf{A}) = |\mathcal{Y}| + R$, hence its columns span $\mathbb{R}^{|\mathcal{Y}|+R}$ and the set $\mathcal{S}$ has an extreme point (see Bertsimas and Tsitsiklis [3, Thm. 2.6]).

We also observe that the objective of (24) is linear in the perception index, hence the maximal value for a given $P$ is attained by some non-increasing linear function of the form $p_0 + p_1 P$. We are now in a position to further develop this insight.

#### A.2.2 Derivation of Eq. (10)

Recall the notation $\rho = \mathbf{D}^\top \mathbf{P}_{X,Y}$. The program (6) can be written as

$$\min_q c^\top q, \text{s.t. } \mathbf{A}q = b, q \ge 0 \tag{28}$$

where $q = [\mathrm{rowstack}\{\mathbf{Q}\}, \mathrm{rowstack}\{\mathbf{\Pi}\}, \varepsilon]$,

$$\mathbf{A} = \left[ \begin{array}{c|c|c} 1^{1\times|\mathcal{X}|} \otimes \mathrm{diag}\{\mathbf{P}_Y\} & 0 & 0 \\ \hline 0 & I_{|\mathcal{X}|} \otimes 1^{1\times|\mathcal{X}|} & 0 \\ \hline I_{|\mathcal{X}|} \otimes \mathbf{P}_Y^\top & -1^{1\times|\mathcal{X}|} \otimes I_{|\mathcal{X}|} & 0 \\ \hline 0 & \mathrm{rowstack}\{H\}^\top & 1 \end{array} \right] \tag{29}$$

and

$$b = \left[ \mathbf{P}_Y^\top, \mathbf{P}_X^\top, 0_{|\mathcal{X}|}, P \right]^\top, \tag{30}$$

$$c = \left[ \mathrm{rowstack}\{\rho\}, 0_{1\times(|\mathcal{X}|^2+1)} \right] \in \mathbb{R}^{|\mathcal{X}|(|\mathcal{Y}|+|\mathcal{X}|)+1}. \tag{31}$$

We now can write (8) as

$$\min_{\tilde{w}} \tilde{w}^\top b, \text{ s.t. } \tilde{w}^\top \mathbf{A} \le c^\top. \tag{32}$$

We denote $\tilde{w} = [w, r, \nu, l]$ where $w \in \mathbb{R}^{|\mathcal{Y}|}$, $r, \nu \in \mathbb{R}^{|\mathcal{X}|}$ and $l$ is a scalar. The first $|\mathcal{X}| \times |\mathcal{Y}|$ columns of $\mathbf{A}$ yield inequalities of the form

$$p_y w_y + p_y \nu_{\hat{x}} \le \rho_{\hat{x},y}, \tag{33}$$

while the next $|\mathcal{X}|^2$ columns yield

$$r_x - \nu_{\hat{x}} + \mathbf{H}_{x,\hat{x}} l \le 0. \tag{34}$$

Finally, the last inequality simply says $l \le 0$.

Now, the above is given in the matrix form

$$\min_{w,r,\nu,l} \left[ w^\top \mathbf{P}_Y + r^\top \mathbf{P}_X + lP \right] \text{ s.t.} \begin{cases} 1^{|\mathcal{X}|\times 1} \otimes (\mathbf{P}_Y \odot w)^\top + 1^{1\times|\mathcal{Y}|} \otimes (\mathbf{P}_Y^\top \odot \nu) \le \rho \\ 1^{|\mathcal{X}|\times 1} \otimes r^\top - 1^{1\times|\mathcal{X}|} \otimes \nu + \mathbf{H}^\top \cdot l \le 0 \\ l \le 0, \end{cases} \tag{35}$$

where $\otimes$ is the Kronecker product and $\odot$ is the Hadamard (elementwise) product. Inequalities between matrices are applied elementwise. To obtain (10), we replace a variable sign ($l \to -l$) and use the connection (9), $\rho = (1^{|\mathcal{X}|} \otimes \mathbf{P}_Y^\top) \odot \rho'$.

It is easy to see that in this case $\mathrm{rank}(\mathbf{A}) = |\mathcal{Y}| + 2|\mathcal{X}|$ while one constraint is redundant, namely we can eliminate a linear constraint from the original program (a row of $\mathbf{A}$) such that the *row rank* of the problem is full. Equivalently, we can set one of the variables $r_x, \nu_{\hat{x}}$ to 0, and the dual feasible set will not contain a line. This implies the existence of an extreme point in the dual set.

## B Piecewise linearity of DP functions (proof of Theorem 4.1)

Here we exploit the properties of the dual problem (24) in order to show that $D(P)$ has a general property - piecewise linearity in the perception index $P$. Moreover, the breakpoints and slopes of this function are determined by the vertices of a convex set in $\mathbb{R}^2$.

We will utilize the following property of LP problems.

**Lemma B.1.** *(Bertsimas and Tsitsiklis [3, Thm. 2.8]) For a bounded LP problem, if there exists an extreme point in the feasible set, then the optimal solution is obtained at an extreme point.*

This is true of course also for the dual problem. We now use this result to prove the following.

**Theorem B.2** (Theorem 4.1 in the main text). *For $P \in [0, \infty)$, the DP function* (6) *is a non-increasing piecewise linear function of $P$ with a non-decreasing slope. Furthermore, there exists $P^* \in [0, 1]$ such that $D(P) = D^*, P \geq P^*$.*

*Proof.* Since it is easier to follow, here we consider the TV index, namely $D(P)$ is given by (2). We emphasize, however, that the same arguments hold for the more general case (6).

Recall $S$ is a matrix whose $j$-th row is $S_j^T$, and $R = 2^{|\mathcal{X}|} - 2$. Let $d = [0^{|\mathcal{Y}|},\ 1^R]^\top$ and $b_0 = [\mathbf{P}_Y^\top,\ -\mathbf{P}_X^\top S^\top]^\top$, both in $\mathbb{R}^{|\mathcal{Y}|+R}$. We can write the objective (24) as

$$[w, \nu]^\top b(P) \rightarrow \max_{w,\nu \in \mathcal{S}}, \tag{36}$$

where $b(P) = b_0 + 2dP$. Let $\text{ext}\,(\mathcal{S}) = \{p^i = [w^i, \nu^i]\}$ denote the vertices of the set of feasible solutions to the dual problem (24), $\mathcal{S}$. Note that the set of vertices is non-empty, finite, and independent of $P$. Lemma B.1 above implies that the dual optimal value is obtained on this set. We now have from strong duality

$$D(P) = \max_i p^i \cdot b(P) = \max_i [w^i, \nu^i]^\top b(P) = \max_i \left[p_0^i + p_1^i P\right], \tag{37}$$

where we denote the projections

$$p_0^i = p^i \cdot b_0, \tag{38}$$

$$p_1^i = p^i \cdot 2d. \tag{39}$$

As a maximum of finite set of linear functions, (37) is a piecewise linear function. The non-decreasing slope property can be easily deduced from (37), or from the fact that DP functions are convex [4]. □

**Corollary B.3.** *The breakpoints of the $D(P)$ function lie within the set*

$$\mathcal{P} = \left\{ \frac{p_0^i - p_0^j}{p_1^j - p_1^i} : \begin{array}{c} p^i, p^j \textit{ are vertices of the set of} \\ \textit{feasible solutions to the dual problem} \end{array} \right\}. \tag{40}$$

As we show next, not every vertex is a candidate for optimality in (37); optimal solutions must be obtained on a 2-D convex hull. Denote the set $\mathcal{S}_2 = \left\{\left(p_0^i, p_1^i\right) : p_0^i = p^i \cdot b_0, p_1^i = p^i \cdot 2d, p^i \in \text{ext}\,(\mathcal{S})\right\} \subseteq \mathbb{R}^2$ which represents the (finite) set of linear curves $\left\{p_0^i + p_1^i P\right\}$ on the 2-dimensional plane.

**Theorem B.4.** *For any $P \geq 0$, there exists a vertex of $\mathcal{S}$ such that $p^k \in \text{argmax}_i\, p^i \cdot b(P)$, and $\left(p_0^k, p_1^k\right)$ is an extreme point of* $\text{conv}\,(\mathcal{S}_2)$.

*Proof.* Let $\left\{\left(\widetilde{p}_0^k, \widetilde{p}_1^k\right)\right\}_{k=1}^M \subseteq \mathcal{S}_2$ be the set of extremals of $\text{conv}\,(\mathcal{S}_2)$. The set $\mathcal{S}_2$ is finite, hence its convex hull is bounded. We can write any point in $\mathcal{S}_2$ as a convex combination $\left(p_0^i, p_1^i\right) = \sum_{k=1}^M \alpha_{ik} \left(\widetilde{p}_0^k, \widetilde{p}_1^k\right)$, thus we have

$$p^i \cdot b(P) = p_0^i + p_1^i P = \sum_{k=1}^M \alpha_{ik} \left(\widetilde{p}_0^k + \widetilde{p}_1^k P\right) \leq \max_k \left(\widetilde{p}_0^k + \widetilde{p}_1^k P\right) = \max_k p^k \cdot b(P). \tag{41}$$

□

## C Full characterization for binary sources (proof of Theorem 4.2)

Recall we discuss the case of binary sources where $\mathcal{X} = \{x_1, x_2\}$ with probabilities $p_{x_1}, p_{x_2}$ respectively, and $\mathcal{Y}$ is of an arbitrary size $n_y$. As a perceptual index, we consider the TV distance (1) while the distortion matrix is arbitrary, yielding the matrix $\rho'$ defined in (9). Denote $u_y = \frac{1}{2}\left(\rho'_{\hat{x}_1 y} - \rho'_{\hat{x}_2 y}\right)$ which is half the cost of reconstructing $y$ as $x_1$ over reconstructing as $x_2$, and we assume *w.l.o.g.* that $u_{y_1} \leq u_{y_2} \leq \ldots \leq u_{y_n}$.We define $P_Y^-(u) = \Pr\{u_Y \leq u\} = \sum_{y: u_y \leq u} \mathbf{P}_Y(y)$. We further denote the symbols $y_i^*$ whose $u_y$ is non-zero, namely

$$u_{-M^-} = u_{y^*_{-M^-}} \leq \ldots \leq u_{-1} = u_{y^*_{-1}} < 0 < u_1 = u_{y^*_1} \leq \ldots \leq u_{M^+} = u_{y^*_{M^+}}. \tag{42}$$

**Theorem C.1.** *(Theorem 4.2 in the main text). Assume that* $p_{x_1} \geq P_Y^-(0)$*, and let* $I = \max\{i\colon p_{x_1} \geq P_Y^-(u_i)\}$*. Then, the DP function* $D(P)$ *is piecewise linear with breakpoints* $\{P_i^*\}_{i=0}^I$ *given by*

$$P_i^* = p_{x_1} - P_Y^-(u_i) \tag{43}$$

*where, specifically,* $P_0^* = p_{x_1} - P_Y^-(0) = P^*$*. The DP function is then given by*

$$D(P) = \begin{cases} D^*, & P \geq P_0^* \\ D(P_{i-1}^*) + 2u_i\left(P_{i-1}^* - P\right), & P_i^* \leq P \leq P_{i-1}^* \\ D(P_I^*) + 2u_{I+1}\left(P_I^* - P\right), & 0 \leq P \leq P_I^* \end{cases}. \tag{44}$$

*If* $P_Y^-(0^-) \geq p_{x_1}$*, then similarly* $P_0^* = P_Y^-(0^-) - p_{x_1}$ *, and* $P_i^* = P_Y^-(u_{-i-1}) - p_{x_1}$*, while it is non-negative, and* $D(P)$ *is determined analogously. In the case* $P_Y^-(0) \geq p_{x_1} \geq P_Y^-(0^-)$*,* $P^* = 0$ *and* $D(P) \equiv D^*$ *for all* $P \geq 0$*.*

*Proof.* Let $\mathcal{X} = \{x_1, x_2\}, \mathcal{Y} = \{y_1, \ldots, y_{n_y}\}$. The dual problem (24) is now written as

$$\begin{cases} \sum_{y\in\mathcal{Y}} p_y w_y & +p_{x_1}(\nu_1 - \nu_2) - p_{x_2}(\nu_1 - \nu_2) - 2P(\nu_1 + \nu_2) \to \max_{w,\nu} \\ \text{s.t.} & \nu_1, \nu_2 \geq 0 \\ & w_y \leq \rho'_{\hat{x}_1 y} - (\nu_1 - \nu_2),\ \rho'_{\hat{x}_2 y} + (\nu_1 - \nu_2), \forall y \in \mathcal{Y} \end{cases}, \tag{45}$$

where we changed the sign of variables $\nu_i \leftarrow -\nu_i$ for convenience. Now, we denote $u = \nu_1 - \nu_2$. Note that $\nu_1 + \nu_2 = |u| + 2\min\{\nu_1, \nu_2\}$. Since $P \geq 0$ and both $\nu_1, \nu_2$ are nonnegative, in an optimal solution we must choose $\min\{\nu_1, \nu_2\}$ to be 0, which implies $|u| = \nu_1 + \nu_2$. The optimization objective in this case boils down to

$$J(u) = \sum_{y\in\mathcal{Y}} p_y \min\left\{\rho'_{\hat{x}_1 y} - u, \rho'_{\hat{x}_2 y} + u\right\} + [2p_{x_1} - 1]\, u - 2P|u|, \tag{46}$$

where in an optimal solution we must have $w_y = \min\left\{\rho'_{\hat{x}_1 y} - u, \rho'_{\hat{x}2y} + u\right\}$ since every $w_y$ should be maximal under the constraints.

We can finally write the Dual objective in this case as

$$\begin{aligned} J_P(u) &= \sum_{y\in\mathcal{Y}} \mathbf{P}_Y(y) \min\{\rho'_{\hat{x}_1 y} - u, \rho'_{\hat{x}_2 y} + u\} + [2p_{x_1} - 1]\, u - 2P|u| & (47) \\ &= \sum_{y: u_y \leq u} \rho_{\hat{x}_1 y} + \sum_{y: u_y > u} \rho_{\hat{x}_2 y} + (1 - P_Y^-(u))u - P_Y^-(u)u + (2p_{x_1} - 1)u - 2P|u| & (48) \\ &= \sum_{y: u_y \leq u} \rho_{\hat{x}_1 y} + \sum_{y: u_y > u} \rho_{\hat{x}_2 y} + 2(p_{x_1} - P_Y^-(u))u - 2P|u|. & (49) \end{aligned}$$

For any $P \geq 0$, this is a *concave* function in the parameter $u$, whose maximal value is obtained on one of the points where the coefficient of $u$ might change its sign.

$$D(P) = \max_u J_P(u) = \max\{J_P(0), J_P(u_{y_1}), \ldots, J_P(u_{y_n})\}. \tag{50}$$

Note that for each $y \in \mathcal{Y}$, $J_P(u_y)$ is a linear function of $P$, with slope $-2|u_y|$. This is true for $u = 0$ as well. The *breakpoints* of $D(P)$ are the points where two (or more) of these functions attain optimality, namely where $\operatorname{argmax}_u J_P(u)$ contains more than one argument. Since $J_P(u)$ is concave w.r.t $u$, $\operatorname{argmax}_u J_P(u)$ must also contain any interval between these points.

As we have already seen, for $P \geq 1$ the DP function is flat,

$$D(P) = D^* = J_P(0) = \sum_y \min_{\hat{x}} \rho_{\hat{x}y}, \quad P \geq 1. \tag{51}$$

It is easy to see from (49) that in fact, $J_P(0) = D^*$ for every $P$. To the right of this point,

$$J_P(u \to 0^+) = D^* + 2(p_{x_1} - P_Y^-(0))u - 2Pu, \tag{52}$$

where to the left,

$$J_P(u \to 0^-) = D^* + 2(p_{x_1} - P_Y^-(0^-))u + 2Pu. \tag{53}$$

We comment that $P_Y^-(0^-) = \sum_{y:u_y<0} \mathbf{P}_Y(y) \leq P_Y^-(0)$, where $P_Y^-(0^-) = P_Y^-(u_{-1})$ if the latter is defined.

Assume now $p_{x_1} \geq P_Y^-(0) \geq P_Y^-(0^-)$, then for every $P \geq 0$, $J_P(u)$ is non-decreasing as $u \to 0^-$. Since it is also concave, the maximal value must be attained at $u = 0$ or on the remaining positive candidate points, which we notate

$$0 < u_1 = u_{y_1^*} \leq \ldots \leq u_{M^+} = u_{y_{M^+}^*}. \tag{54}$$

At the first breakpoint $P_0^* = P^*$, where $0, u_1 \in \operatorname{argmax} J_P(u)$, we should have $J_{P^*}(0) = J_{P^*}(u_1)$, or equivalently

$$2P_0^* u = 2(p_{x_1} - P_Y^-(0))u, \; 0 \leq u < u_1, \tag{55}$$

yielding $P_0^* = p_{x_1} - P_Y^-(0)$. Similarly, for every possible breakpoint we should have

$$2P_i^* u = 2(p_{x_1} - P_Y^-(u_i))u, \; u_i \leq u < u_{i+1}, \tag{56}$$

implying $P_i^* = p_{x_1} - P_Y^-(u_i)$ for every $i$ such that $p_{x_1} \geq P_Y^-(u_i)$. (for a discussion about the case where $u_i$ might be equal to $u_{i+1}$ we refer the reader to Remark 4.3 in the main text).

If $P_Y^-(0) \geq P_Y^-(0^-) \geq p_{x_1}$, then (49) is non-decreasing as $u \to 0^+$. Now, maximum must occur at $u = 0$ or the remaining *negative* candidates, $u_{-i}$. By arguments similar to the case above, $P_0^* = P_Y^-(0^-) - p_{x_1}$, and $P_i^* = P_Y^-(u_{-i}^-) - p_{x_1} = P_Y^-(u_{-i-1}) - p_{x_1}$ while it is non-negative.

Finally, in the case $P_Y^-(0) \geq p_{x_1} \geq P_Y^-(0^-)$, (53) is non-decreasing, while (52) is non-increasing for every $P$, implying $max_u J_P(u) = J_P(0) = D^*$ hence in this case $P^* = 0, D(P) \equiv D^*$. □