A Theory of the Distortion-Perception Tradeoff in Wasserstein Space

Anonymous Author(s)
Affiliation
Address
email

Abstract

The lower the distortion of an estimator, the more the distribution of its outputs generally deviates from the distribution of the signals it attempts to estimate. This phenomenon, known as the perception-distortion tradeoff, has captured significant attention in image restoration, where it implies that fidelity to ground truth images comes at the expense of perceptual quality (deviation from statistics of natural images). However, despite the increasing popularity of performing comparisons on the perception-distortion plane, there remains an important open question: *what is the minimal distortion that can be achieved under a given perception constraint?* In this paper, we derive a closed form expression for this distortion-perception (DP) function for the mean squared-error (MSE) distortion and the Wasserstein-2 perception index. We prove that the DP function is always quadratic, regardless of the underlying distribution. This stems from the fact that estimators on the DP curve form a geodesic in Wasserstein space. In the Gaussian setting, we further provide a closed form expression for such estimators. For general distributions, we show how these estimators can be constructed from the estimators at the two extremes of the tradeoff: The global MSE minimizer, and a minimizer of the MSE under a perfect perceptual quality constraint. The latter can be obtained as a stochastic transformation of the former.

1 Introduction

Image restoration covers some fundamental settings in image processing such as denoising, deblurring and super-resolution. Over the past few years, image restoration methods have demonstrated impressive improvements in both visual quality and distortion measures such as peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [31]. It was noticed, however, that improvement in accuracy, as measured by distortion, does not necessarily lead to improvement in visual quality, referred to as perceptual quality. Furthermore, the lower the distortion of an estimator, the more the distribution of its outputs generally deviates from the distribution of the signals it attempts to estimate. This phenomenon, known as the *perception-distortion tradeoff* [4], has captured significant attention, where it implies that faithfulness to ground truth images comes at the expense of perceptual quality, namely the deviation from statistics of natural images. Several works have extended the perception-distortion tradeoff to settings such as lossy compression [5] and classification [14].

Despite the increasing popularity of performing comparisons on the perception-distortion plane, the exact characterization of the minimal distortion that can be achieved under a given perception constraint remains an important open question. Although Blau and Michaeli [4] investigated the basic properties of this *distortion-perception function*, such as monotonicity and convexity, little is known about its precise nature. While a general answer to this question is unavailable, in this

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

paper, we derive a closed form expression for the distortion-perception (DP) function for the mean squared-error (MSE) distortion and the Wasserstein-2 perception index.

Our main contributions are: *(i)* We prove that the DP function is *always* quadratic in the perception constraint P, regardless of the underlying distribution (Theorem 1). *(ii)* We show that it is possible to construct estimators on the DP curve from the estimators at the two extremes of the tradeoff (Theorem 3): The one that globally minimizes the MSE, and a minimizer of the MSE under a perfect perceptual quality constraint. The latter can be obtained as a stochastic transformation of the former. *(iii)* In the Gaussian setting, we further provide a closed form expression for optimal estimators and for the corresponding DP curve (Theorems 4 and 5). We show this Gaussian DP curve is a lower bound on the DP curve of any distribution having the same second order statistics. Finally, we illustrate our results, numerically and visually, in a super-resolution setting in Section 5. The proofs of all the theorems in the main text are provided in Appendix B

Our theoretical results shed light on several topics that are subject to much practical activity. Particularly, in the domain of image restoration, numerous works target perceptual quality rather than distortion (*e.g.* [29, 13, 12]). However, it has recently been recognized that generating a single reconstructed image often does not convey to the user the inherent ambiguity in the problem. Therefore, many recent works target *diverse* perceptual image reconstruction, by employing randomization among possible restorations [15, 3, 21, 1]. Commonly, such works perform sampling from the posterior distribution of natural images given the degraded input image. This is done *e.g.* using priors over image patches [7], conditional generative models [18, 20], or implicit priors induced by deep denoiser networks [10]. Theoretically, posterior sampling leads to perfect perceptual quality (the restored outputs are distributed like the prior). However, a fundamental question is whether this is optimal in terms of distortion. As we show in Section 3.1, posterior sampling is often not an optimal strategy, in the sense that there often exist perfect perceptual quality estimators that achieve lower distortion.

Another topic of practical interest, is the ability to *traverse the distortion-perception tradeoff* at test time, without having to train a different model for each working point. Recently, interpolation has been suggested for controlling several objectives at test-time. Shoshan et al. [25] propose using interpolation in some latent space in order to approximate intermediate objectives. Wang et al. [29] use per-pixel interpolation for balancing perceptual quality and fidelity. Studies of network parameter interpolation are presented by Wang et al. [29, 30]. Deng [6] produces a low distortion reconstruction and a high perceptual quality one, and then uses style transfer to combine them. An important question, therefore, is which strategy is optimal. In Section 3.2 we show that for the MSE–Wasserstein-2 tradeoff, linear interpolation leads to optimal estimators. We also discuss a geometric connection between interpolation and the fact that estimators on the DP curve form a geodesic in Wasserstein space.

2 Problem setting and preliminaries

2.1 The distortion-perception tradeoff

Let X, Y be random vectors taking values in $\mathbb{R}^{n_x}$ and $\mathbb{R}^{n_y}$, respectively. We consider the problem of constructing an estimator $\hat{X}$ of X based on Y. Namely, we are interested in determining a conditional distribution $p_{\hat{X}|Y}$ such that $\hat{X}$ constitutes a good estimate of X.

In many practical cases, the goodness of an estimator is associated with two factors: (i) the degree to which $\hat{X}$ is close to X on average (low distortion), and (ii) the degree to which the distribution of $\hat{X}$ is close to that of X (good perceptual quality). An important question, then, is *what is the minimal distortion that can be achieved under a given level of perceptual quality?* and *how can we construct estimators that achieve this lower bound?* In mathematical language, we are interested in analyzing the distortion-perception (DP) function (defined similarly to the perception-distortion function of [4])

$$D(P) = \min_{p_{\hat{X}|Y}} \left\{ \mathbb{E}[d(X, \hat{X})] \ : \ d_p(p_X, p_{\hat{X}}) \leq P \right\}. \tag{1}$$

Here, $d : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^+ \cup \{0\}$ is some distortion criterion, $d_p(\cdot, \cdot)$ is some divergence between probability measures, and $p_{\hat{X}}$ is the probability measure on $\mathbb{R}^{n_x}$ induced by $p_{\hat{X}|Y}$ and p_Y. We assume that $\hat{X}$ is independent of X given Y.

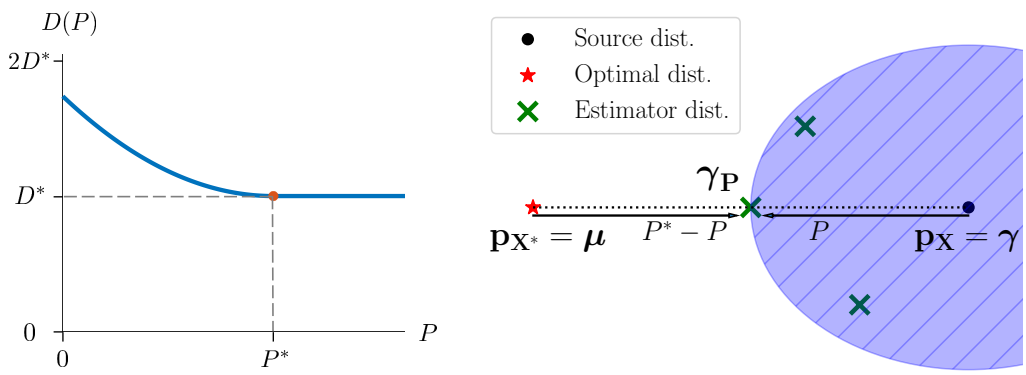

Figure 1: **Left: The distortion-perception function.** When using the MSE distortion and the Wasserstein-2 perception index, the minimal possible distortion, D^*, is achieved by the estimator $X^* = E[X|Y]$. The perception index attained by this estimator is P^*. At the other extreme of the tradeoff, we know that the distortion at $P = 0$ is bounded from above by $2D^*$. **Right:** The minimal distortion $D(P)$ for a given perception index $P < P^*$ can be achieved by an estimator with a distribution γ_P lying on a straight line (or geodesic) defined by the geometry of the probabilities space. Given P, γ_P achieves $W_2(p_X, \gamma_P) = P$ and $W_2(p_{X^*}, \gamma_P) = P^* - P$, hence $D(P) = D^* + W_2^2(p_{X^*}, \gamma_P) = D^* + (P^* - P)^2$.

As discussed in [4], the function $D(P)$ is monotonically non-increasing and is convex whenever $d_p(\cdot,\cdot)$ is convex in its second argument (which is the case for most popular divergences). However, without further concrete assumptions on the distortion measure $d(\cdot,\cdot)$ and the perception index $d_p(\cdot,\cdot)$, little can be said about the precise nature of $D(P)$.

Here, we focus our attention on the squared-error distortion $d(x,\hat{x}) = \|x-\hat{x}\|^2$ and the Wasserstein-2 distance $d_p(p_X, p_{\hat{X}}) = W_2(p_X, p_{\hat{X}})$, with which (1) reads

$$D(P) = \min_{p_{\hat{X}|Y}} \left\{ \mathbb{E}[\|X - \hat{X}\|^2] \ : \ W_2(p_X, p_{\hat{X}}) \leq P \right\}. \tag{2}$$

Throughout this paper we assume that all distributions have finite first and second moments. In addition, from Theorem 3 below it will follow that the minimum is indeed attained, so that (2) is well defined. It is well known that the estimator minimizing the mean squared error (MSE) without any constraints, is given by $X^* = \mathbb{E}[X|Y]$. This implies that $D(P)$ monotonically decreases until P reaches $P^* \triangleq W_2(p_X, p_{X^*})$, beyond which point $D(P)$ takes the constant value $D^* \triangleq \mathbb{E}[\|X - X^*\|^2]$. This is illustrated in Fig. 1. It is also known that in this case $D(0) \leq 2D^*$ since the posterior sampling estimator $p_{\hat{X}|Y} = p_{X|Y}$ achieves $W_2(p_X, p_{\hat{X}}) = 0$ and $\mathbb{E}[\|X - \hat{X}\|^2] = 2D^*$ [4]. However, apart for these rather general properties, the precise shape of the DP curve has not been determined to date, and neither have the estimators that achieve the optimum in (2). This is our goal in this paper.

2.2 The Wasserstein and Gelbrich Distances

Before we present our main results, we briefly survey a few properties of the Wasserstein distance, mostly taken from [19]. The Wasserstein-p ($p \geq 1$) distance between measures μ and γ on a separable Banach space $\mathcal{X}$ with norm $\|\cdot\|$ is defined by

$$W_p^p(\mu, \gamma) \triangleq \inf \left\{ \mathbb{E}_{(U,V)\sim\nu}[\|U - V\|^p] \ : \ \nu \in \Pi(\mu, \gamma) \right\}, \tag{3}$$

where $\Pi(\mu, \gamma)$ is the set of all probabilities on $\mathcal{X} \times \mathcal{X}$ with marginals μ and γ. A joint probability ν achieving the optimum in (3) is often referred to as *optimal plan*. The Wasserstein space of probability measures is defined as

$$\mathcal{W}_p(\mathcal{X}) \triangleq \left\{ \gamma : \int_{\mathcal{X}} \|x\|^p d\gamma < \infty \right\},$$

and W_p constitutes a metric on $\mathcal{W}_p(\mathcal{X})$.

For any $(m_1, \Sigma_1), (m_2, \Sigma_2) \in \mathbb{R}^d \times \mathbb{S}_+^d$ (where $\mathbb{S}_+^d$ is the set of symmetric positive semidefinite matrices in $\mathbb{R}^{d\times d}$), the Gelbrich distance is defined as

$$G^2((m_1, \Sigma_1), (m_2, \Sigma_2)) \triangleq \|m_1 - m_2\|_2^2 + \mathrm{Tr}\left\{\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}}\right)^{\frac{1}{2}}\right\}. \tag{4}$$

The root of a PSD matrix is always taken to be PSD. For any two probability measures μ_1, μ_2 on $\mathbb{R}^d$ with means and covariances $(m_1, \Sigma_1), (m_2, \Sigma_2)$, from [8, Thm. 2.1] we have that

$$W_2^2(\mu_1, \mu_2) \geq G^2((m_1, \Sigma_1), (m_2, \Sigma_2)). \tag{5}$$

When $\mu_1 = \mathcal{N}(m_1, \Sigma_1)$ and $\mu_2 = \mathcal{N}(m_2, \Sigma_2)$ are Gaussian distributions on $\mathbb{R}^d$, we have that $W_2(\mu_1, \mu_2) = G((m_1, \Sigma_1), (m_2, \Sigma_2))$. This equality is obvious for non-singular measures but is true for any two Gaussian distributions [19, p. 18]. If Σ_1 and Σ_2 are non-singular, then the distribution attaining the optimum in (3) corresponds to

$$U \sim \mathcal{N}(m_1, \Sigma_1), \quad V = m_2 + T_{1\to 2}(U - m_1), \tag{6}$$

where

$$T_{1\to 2} = \Sigma_1^{-\frac{1}{2}}\left(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}}\right)^{\frac{1}{2}}\Sigma_1^{-\frac{1}{2}} \tag{7}$$

is the optimal transformation pushing forward from $\mathcal{N}(0, \Sigma_1)$ to $\mathcal{N}(0, \Sigma_2)$ [11]. This transformation satisfies $\Sigma_2 = T_{1\to 2}\Sigma_1 T_{1\to 2}$. For a discussion on singular distributions, please see App. A.

3 Main results

3.1 The MSE–Wasserstein-2 tradeoff

The DP function (2) depends, of course, on the underlying joint probability p_{XY} of the signal X and measurements Y. Our first key result is that this dependence can be expressed solely in terms of D^* and P^*. In other words, knowing the distortion and perception index attained by the minimum MSE estimator X^*, suffices for determining $D(P)$ for any P.

Theorem 1 (The DP function)**.** *The DP function* (2) *is given by*

$$D(P) = D^* + \left[(P^* - P)_+\right]^2, \tag{8}$$

where $(x)_+ = \max(0, x)$*. Furthermore, an estimator achieving perception index* P *and distortion* $D(P)$ *can always be constructed by applying a (possibly stochastic) transformation to* X^**.*

Theorem 1 is of practical importance because in many cases constructing an estimator that achieves a low MSE (*i.e.* an approximation of X^*) is a rather simple task. This is the case, for example, in image restoration with deep neural networks. There, it is common practice to train a network by minimizing its average squared error on a training set. Now, measuring the MSE of such a network on a large test set allows approximating D^*. We can also obtain an approximation of at least a lower bound on P^* by estimating the second order statistics of X and X^*. Specifically, recall that P^* is lower bounded by the Gelbrich distance between (m_X, Σ_X) and (m_{X^*}, Σ_{X^*}), which is given by $(G^*)^2 \triangleq \mathrm{Tr}\{\Sigma_X + \Sigma_{X^*} - 2(\Sigma_X^{1/2}\Sigma_{X^*}\Sigma_X^{1/2})^{1/2}\}$ (see (5)). Given approximations for D^* and G^*, we can approximate a lower bound on the DP function for any P,

$$D(P) \geq D^* + [(G^* - P)_+]^2. \tag{9}$$

The bound is attained when X and Y are jointly Gaussian.

Uniqueness A remark is in place regarding the uniqueness of an estimator achieving (8). As we discuss below, what defines an optimal estimator $\hat{X}$ is its joint distribution with X^*. This joint distribution may not be unique, in which case the optimal estimator is not unique. Moreover, even if $p_{\hat{X}X^*}$ is unique, the uniqueness of the estimator is not guaranteed because there may be different conditional distributions $p_{\hat{X}|Y}$ that lead to the same optimal $p_{\hat{X}X^*}$. In other words, given the optimal $p_{\hat{X}X^*}$, one can choose any joint probability $p_{\hat{X}YX^*}$ that has marginals $p_{\hat{X}X^*}$ and p_{YX^*}. One option is to take the estimator $\hat{X}$ to be a (possibly stochastic) transformation of X^*, namely $p_{\hat{X}|Y} = p_{\hat{X}|X^*}p_{X^*|Y}$. But there may be other options. In cases where either Y or $\hat{X}$ are a deterministic transformation of X^* (*e.g.* when X^* has a density, or is an invertible function of Y), there is a unique joint distribution $p_{\hat{X}YX^*}$ with the given marginals [2, Lemma 5.3.2]. In this case, if $p_{\hat{X}X^*}$ is unique then so is the estimator $p_{\hat{X}|Y}$.

Randomness Under the settings of image restoration, many methods encourage diversity in their outpus by adding randomness [15, 3, 21]. In our setting, we may ask under what conditions there exists an optimal estimator $\hat{X}$ which is a deterministic function of Y. For example, when $p_Y = \delta_0$ but X has some non-atomic distribution, it is clear that no deterministic function of Y can attain perfect perceptual quality. It turns out that a sufficient condition for the optimal $\hat{X}$ to be a deterministic function of Y is that X^* have a density. We discuss this in App. B and explicitly illustrate it in the Gaussian case (see Sec. 3.3), where if X^* has a non-singular covariance matrix then $\hat{X}$ is a deterministic function of Y.

When is posterior sampling optimal? Many recent image restoration methods attempt to produce diverse high perceptual quality reconstructions by sampling from the posterior distribution $p_{X|Y}$ [7, 18, 10]. As discussed in [4], the posterior sampling estimator attains a perception index of 0 (namely $W_2(p_X, p_{\hat{X}}) = 0$) and distortion $2D^*$. But an interesting question is: when is this strategy optimal? In other words, in what cases do we have that the DP function at $P = 0$ equals precisely $2D^*$ and is not strictly smaller? Note from the definition of the Wasserstein distance (3), that $(P^*)^2 = W_2^2(p_X, p_{X^*}) \leq \mathbb{E}[\|X - X^*\|^2] = D^*$. Using this in (8) shows that the DP function at $P = 0$ is upper bounded by

$$D(0) = D^* + (P^*)^2 \leq 2D^*, \tag{10}$$

and the upper bound is attained when $(P^*)^2 = D^*$. To see when this happens, observe that

$$\mathrm{Tr}\left\{\Sigma_X + \Sigma_{X^*} - 2(\Sigma_X^{\frac{1}{2}}\Sigma_{X^*}\Sigma_X^{\frac{1}{2}})^{\frac{1}{2}}\right\} = (G^*)^2 \leq (P^*)^2 \leq D^* = \mathrm{Tr}\{\Sigma_X - \Sigma_{X^*}\}. \tag{11}$$

We can see that when $\mathrm{Tr}\{\Sigma_{X^*}\} = \mathrm{Tr}\{(\Sigma_X^{1/2}\Sigma_{X^*}\Sigma_X^{1/2})^{1/2}\}$, the leftmost and rightmost sides become equal, and thus $(P^*)^2 = D^*$. To understand the meaning of this condition, let us focus on the case where Σ_X and Σ_{X^*} are jointly diagonalizable. This is a reasonable assumption for natural images, where shift-invariance induces diagonalization by the Fourier basis [28]. In this case, the condition can be written in terms of the eigenvalues of the matrices, namely $\sum_i \lambda_i(\Sigma_{X^*}) = \sum_i \sqrt{\lambda_i(\Sigma_{X^*})\lambda_i(\Sigma_X)}$. This condition is satisfied when each $\lambda_i(\Sigma_{X^*})$ equals either $\lambda_i(\Sigma_X)$ or 0. Namely, the ith eigenvalue of the error covariance of X^*, which is given by $\Sigma_X - \Sigma_{X^*}$, is either $\lambda_i(\Sigma_X)$ or 0. We conclude that posterior sampling is optimal when there exists a subspace $\mathcal{S}$ spanned by some of the eigenvectors of Σ_X, such that the projection of X onto $\mathcal{S}$ can be recovered from Y with zero error, but the projection of X onto $\mathcal{S}^\perp$ cannot be recovered at all (the optimal estimator is trivial). This is likely not the case in most practical scenarios. Therefore, it seems that *posterior sampling is often not optimal*. That is, posterior sampling can be improved upon in terms of MSE without any sacrifice in perceptual quality.

3.2 Optimal estimators

While Theorem 1 reveals the shape of the DP function, it does not provide a recipe for constructing optimal estimators on the DP tradeoff. We now discuss the nature of such estimators.

Our first observation is that since $\hat{X}$ is independent of X given Y, its MSE can be decomposed as $\mathbb{E}[\|X - \hat{X}\|^2] = \mathbb{E}[\|X - X^*\|^2 + \mathbb{E}[\|X^* - \hat{X}\|^2]$ (see App. B). Therefore, the DP function (2) can be equivalently written as

$$D(P) = D^* + \min_{p_{\hat{X}|Y}} \left\{\mathbb{E}[\|\hat{X} - X^*\|^2] \ : \ W_2(p_X, p_{\hat{X}}) \leq P\right\}. \tag{12}$$

Note that the objective in (12) depends on the MSE between $\hat{X}$ and X^*, so that we can perform the minimization on $p_{\hat{X}|X^*}$ rather than on $p_{\hat{X}|Y}$ (once we determine the optimal $p_{\hat{X}|X^*}$ we can construct a consistent $p_{\hat{X}|Y}$ as discussed above).

Now, let us start by examining the leftmost side of the curve $D(P)$, which corresponds to a perfect perceptual quality estimator (*i.e.* $P = 0$). In this case, the constraint becomes $p_{\hat{X}} = p_X$. Therefore,

$$D(0) = D^* + \min_{p_{\hat{X}X^*}} \left\{\mathbb{E}[\|\hat{X} - X^*\|^2] \ : \ p_{\hat{X}X^*} \in \Pi(p_X, p_{X^*})\right\}, \tag{13}$$

where $\Pi(p_X, p_{X^*})$ is the set of all probabilities on $\mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ with marginals p_X, p_{X^*}. One may readily recognize this as the optimization problem underlying the Wasserstein-2 distance between p_X and p_{X^*}. This leads us to the following conclusion.

Theorem 2 (Optimal estimator for $P = 0$)**.** *Let $\hat{X}_0$ be an estimator achieving perception index 0 and MSE $D(0)$. Then its joint distribution with X^* attains the optimum in the definition of $W_2(p_X, p_{X^*})$. Namely, $p_{\hat{X}_0 X^*}$ is an optimal plan between p_X and p_{X^*}.*

Having understood the estimator $\hat{X}_0$ at the leftmost end of the tradeoff, we now turn to study optimal estimators for arbitrary P. Interestingly, we can show that Problem (12) is equivalent to (see App. B)

$$D(P) = D^* + \min_{p_{\hat{X}}} \left\{ W_2^2(p_{\hat{X}}, p_{X^*}) \ : \ W_2(p_{\hat{X}}, p_X) \leq P \right\}. \tag{14}$$

Namely, an optimal $p_{\hat{X}}$ is closest to p_{X^*} among all distributions within a ball of radius P around p_X, as illustrated in Fig. 1. Moreover, $p_{\hat{X}X^*}$ is an optimal plan between $p_{\hat{X}}$ and p_{X^*}. As it turns out, this somewhat abstract viewpoint leads to a rather practical construction for $\hat{X}$ from the estimators $\hat{X}_0$ and X^* at the two extremes of the tradeoff. Specifically, we have the following result, proved in App. B.

Theorem 3 (Optimal estimators for arbitrary P)**.** *Let $\hat{X}_0$ be an estimator achieving perception index 0 and MSE $D(0)$. Then for any $P \in [0, P^*]$, the estimator*

$$\hat{X}_P = \left(1 - \frac{P}{P^*}\right) \hat{X}_0 + \frac{P}{P^*} X^* \tag{15}$$

is optimal for perception index P. Namely, it achieves perception index P and distortion $D(P)$.

Theorem 3 has important implications for perceptual signal restoration. For example, in the task of image super-resolution, there exist many deep network based methods that achieve a low MSE [13, 27, 24]. These provide an approximation for X^*. Moreover, there is an abundance of methods that achieve good perceptual quality at the price of a reasonable degradation in MSE (often by incorporating a GAN-based loss) [12, 29, 23]. These constitute approximations for $\hat{X}_0$. However, achieving results that strike other prescribed balances between MSE and perceptual quality commonly require training a different model for each setting. Shoshan et al. [25] and Navarrete Michelini et al. [17] tried to address this difficulty by introducing new training techniques that allow traversing the distortion-perception tradeoff at test time. But, interestingly, Theorem 3 shows that in our setting such specialized training methods are not required. Having a model that leads to low MSE and one that leads to good perceptual quality, it is possible to construct any other estimator on the DP tradeoff, by simply averaging the outputs of these two models with appropriate weights. We illustrate this in Sec. 5.

3.3 The Gaussian setting

When X and Y are jointly Gaussian, it is well known that the minimum MSE estimator X^* is a linear function of the measurements Y. However, it is not *a-priori* clear whether all estimators along the DP tradeoff are linear in this case, and what kind of randomness they possess. As we now show, equipped with Theorem 3, we can obtain closed form expressions for optimal estimators for any P. For simplicity, we assume here that X and Y have zero means and that $\Sigma_X, \Sigma_Y \succ 0$.

It is instructive to start by considering the simple case, where Σ_{X^*} is non-singular (in Theorem 4 below we address the more general case of a possibly singular Σ_{X^*}). It is well known that

$$X^* = \Sigma_{XY} \Sigma_Y^{-1} Y, \qquad \Sigma_{X^*} = \Sigma_{XY} \Sigma_Y^{-1} \Sigma_{YX}. \tag{16}$$

Now, since we assumed that $\Sigma_X, \Sigma_{X^*} \succ 0$, we have from Theorem 2 and (6),(7) that

$$\hat{X}_0 = \Sigma_{X^*}^{-\frac{1}{2}} \left(\Sigma_{X^*}^{\frac{1}{2}} \Sigma_X \Sigma_{X^*}^{\frac{1}{2}} \right)^{\frac{1}{2}} \Sigma_{X^*}^{-\frac{1}{2}} X^*. \tag{17}$$

Finally, we know that $P^* = G^*$, which is given by the left-hand side of (11). Substituting these expressions into (15), we obtain that an optimal estimator for perception $P \in [0, G^*]$ is given by

$$\hat{X}_P = \left(\left(1 - \frac{P}{G^*}\right) \Sigma_{X^*}^{-\frac{1}{2}} \left(\Sigma_{X^*}^{\frac{1}{2}} \Sigma_X \Sigma_{X^*}^{\frac{1}{2}} \right)^{\frac{1}{2}} \Sigma_{X^*}^{-\frac{1}{2}} + \frac{P}{G^*} I \right) \Sigma_{XY} \Sigma_Y^{-1} Y. \tag{18}$$

As can be seen, this optimal estimator is a deterministic linear transformation of Y for any P.

The setting just described does not cover the case where Y is of lower dimensionality than X because in that case Σ_{X^*} is necessarily singular (it is a $N_x \times N_x$ matrix of rank at most N_y; see (16)). In this case, any deterministic linear function of Y would result in an estimator $\hat{X}$ with a rank-N_y covariance. Obviously, the distribution of such an estimator cannot be arbitrarily close to that of X, whose covariance has rank N_x. What is the optimal estimator in this more general setting, then?

Theorem 4 (Optimal estimators in the Gaussian case)**.** *Assume X and Y are zero-mean jointly Gaussian random vectors with $\Sigma_X, \Sigma_Y \succ 0$. Denote $T^* \triangleq T_{p_X \to p_{X^*}} = \Sigma_X^{-1/2}(\Sigma_X^{1/2}\Sigma_{X^*}\Sigma_X^{1/2})^{1/2}\Sigma_X^{-1/2}$. Then for any $P \in [0, G^*]$, an estimator with perception index P and MSE $D(P)$ can be constructed as*

$$\hat{X}_P = \left(\left(1 - \frac{P}{G^*}\right)\Sigma_X^{\frac{1}{2}}\left(\Sigma_X^{\frac{1}{2}}\Sigma_{X^*}\Sigma_X^{\frac{1}{2}}\right)^{\frac{1}{2}}\Sigma_X^{-\frac{1}{2}}\Sigma_{X^*}^{\dagger} + \frac{P}{G^*}I\right)\Sigma_{XY}\Sigma_Y^{-1}Y + \left(1 - \frac{P}{G^*}\right)W, \quad (19)$$

where W is a zero-mean Gaussian noise with covariance $\Sigma_W = \Sigma_X^{1/2}(I - \Sigma_X^{1/2}T^*\Sigma_{X^*}^{\dagger}T^*\Sigma_X^{1/2})\Sigma_X^{1/2}$, which is independent of Y, X, and $\Sigma_{X^*}^{\dagger}$ is the pseudo-inverse of Σ_{X^*}.

Note that in this case, we indeed have a random noise component that shapes the covariance of $\hat{X}_P$ to become closer to Σ_X as P gets closer to 0. It can be shown (see App. B) that when Σ_{X^*} is invertible, $\Sigma_W = 0$ and (19) reduces to (18). Also note that, as in (18), the dependence of $\hat{X}_P$ on Y in (19) is only through $X^* = \Sigma_{XY}\Sigma_Y^{-1}Y$.

As mentioned in Sec. 3.1, the optimal estimator is generally not unique. Interestingly, in the Gaussian setting we can explicitly characterize a *set* of optimal estimators.

Theorem 5 (A set of optimal estimators in the Gaussian case)**.** *Consider the setting of Theorem 4. Let $\Sigma_{\hat{X}_0 Y} \in \mathbb{R}^{n_x \times n_y}$ satisfy*

$$\Sigma_{\hat{X}_0 Y}\Sigma_Y^{-1}\Sigma_{YX} = \Sigma_X^{\frac{1}{2}}(\Sigma_X^{\frac{1}{2}}\Sigma_{X^*}\Sigma_X^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_X^{-\frac{1}{2}}, \quad (20)$$

and W_0 be a zero-mean Gaussian noise with covariance

$$\Sigma_{W_0} = \Sigma_X - \Sigma_{\hat{X}_0 Y}\Sigma_Y^{-1}\Sigma_{\hat{X}_0 Y}^T \succeq 0 \quad (21)$$

that is independent of X, Y. Then, for any $P \in [0, G^*]$, an optimal estimator with perception index P can be obtained by

$$\hat{X}_P = \left(\left(1 - \frac{P}{G^*}\right)\Sigma_{\hat{X}_0 Y} + \frac{P}{G^*}\Sigma_{XY}\right)\Sigma_Y^{-1}Y + \left(1 - \frac{P}{G^*}\right)W_0. \quad (22)$$

The estimator given in (19) *is one solution to* (20)-(21)*, but is generally not unique.*

4 A geometric perspective on the distortion-perception tradeoff

In this section we provide a geometric point of view on our main results. Specifically, we show that the results of Theorems 1 and 3 are a consequence of a more general geometric property of the space $\mathcal{W}_2(R^{n_x})$. In the Gaussian case, this is simplified to a geometry of covariance matrices.

Recall from (14) that the optimal $p_{\hat{X}}$ is the one closest to p_{X^*} (in terms of Wasserstein distance) among all measures at a distance P from p_X. This implies that to determine $p_{\hat{X}}$, we should traverse the geodesic between p_{X^*} and p_X until reaching a distance of P from p_X. Furthermore, $p_{\hat{X}X^*}$ should be the optimal plan between $p_{\hat{X}}$ and p_{X^*}. Interestingly, geodesics in Wasserstein spaces take a particularly simple form, and their explicit construction also turns out to satisfy the latter requirement. Specifically, let γ, μ be measures in $\mathcal{W}_2(R^d)$, let $\nu \in \Pi(\gamma, \mu)$ be an optimal plan attaining $W_2(\gamma, \mu)$, and let π_i denote the projection $\pi_i : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ such that $\pi_i((x_1, x_2)) = x_i,\ i = 1, 2$. Then, the curve

$$\gamma_t \triangleq [(1-t)\pi_1 + t\pi_2]\,\#\nu, \quad t \in [0, 1] \quad (23)$$

is a constant-speed geodesic from γ to μ in $\mathcal{W}_2(\mathbb{R}^d)$ [2], where $\#$ is the push-forward operation[1]. Particularly,

$$W_2(\gamma_t, \gamma_s) = |t - s| W_2(\gamma, \mu), \quad (24)$$

[1]For measures γ, μ on $\mathcal{X}, \mathcal{Y}$, we say that a measurable transform $T : \mathcal{X} \to \mathcal{Y}$ pushes γ forward to μ (denoted $T\#\gamma = \mu$) iff $\gamma(T^{-1}(B)) = \mu(B)$ for any measurable $B \subseteq \mathcal{Y}$.

and it follows that $W_2(\gamma_t, \gamma) = tW_2(\gamma, \mu)$ and $W_2(\gamma_t, \mu) = (1-t)W_2(\gamma, \mu)$. Furthermore, if $\gamma_t, t \in [0,1]$ is a constant-speed geodesic with $\gamma_0 = \gamma, \gamma_1 = \mu$, then the optimal plans between γ, γ_t and between γ_t, μ are given by

$$[\pi_1, (1-t)\pi_1 + t\pi_2] \#\nu, \quad [(1-t)\pi_1 + t\pi_2, \pi_2] \#\nu, \tag{25}$$

respectively, where $\nu \in \Pi(\gamma, \mu)$ is some optimal plan. Applying (23) to $(\hat{X}_0, X^*) \sim \nu$ with $t = P/P^*$, we obtain (15), where we show that the obtained estimator achieves $\mathbb{E}[\|\hat{X}_P - X^*\|^2] = (1-t)^2 W_2^2(p_X, p_{X^*})$. This explains the result of Theorem 3.

It is worth mentioning that this geometric interpretation is simplified under some common settings. For example, when γ is absolutely continuous (w.r.t. the Lebesgue measure), we have a measurable map $T_{\gamma\to\mu}$ which is the solution to the optimal transport problem with the quadratic cost (Panaretos and Zemel [19], Thm 1.6.2, p.16). The geodesic (23) then takes the form

$$\gamma_t = [Id + t(T_{\gamma\to\mu} - Id)]\#\gamma, \quad t \in [0,1]. \tag{26}$$

Therefore, in our setting, if $\gamma = p_{X^*}$ has a density, then we can obtain $\hat{X}_P$ by the deterministic transformation $[X^* + \left(1 - \frac{P}{P^*}\right)(T_{p_{X^*}\to p_X}(X^*) - X^*)]$ (see Remark about randomness in Sec. 3.1).

Further simplification arises when γ, μ are centered non-singular Gaussian measures, in which case $T_{\gamma\to\mu}$ is the linear and symmetric transformation (7). Then, γ_t is a Gaussian measure with covariance $\Sigma_{\gamma_t} = T_t \Sigma_\gamma T_t$, where $T_t \triangleq [I + t(T_{\gamma\to\mu} - I)]$. Therefore, in the Gaussian case, the shortest path (23) between distributions is reduced to a trajectory in the geometry of covariance matrices induced by the Gelbrich distance [26]. If additionally Σ_γ and Σ_μ commute, then the Gelbrich distance is further reduced to the ℓ^2-distance between matrices, as we discuss in App. C.

5 Numerical illustration

In this Section we evaluate 12 super resolution algorithms on the BSD100 dataset[2] [16]. The evaluated algorithms include EDSR [13], ESRGAN [29], SinGAN [23], ZSSR [24], DIP [27], SRResNet variants which optimize MSE and $\text{VGG}_{2,2}$, SRGAN variants which optimize MSE, $\text{VGG}_{2,2}$ and $\text{VGG}_{5,4}$ in addition to an adversarial loss [12], ENet [22] ("PAT" and "E" variants). Low resolution images were obtained by $4\times$ downsampling using a bicubic kernel.

In Figure 2 we plot each method on the distortion-perception plane. Specifically, we consider natural (and reconstructed) images to be stationary random sources, and use 9×9 patches (totally 1.6×10^6 patches) from the RGB images to empirically estimate the mean and covariance matrix for the ground-truth images, and for the reconstructions produced by each method. We then use the estimated Gelbrich distances (4) between the patch distribution of each method and that of ground-truth images, as a perceptual quality index. Recall this is a lower bound on the Wasserstein distance.

We consider EDSR [13] to be the best MSE estimator X^* since it achieves the lowest distortion among the evaluated methods. We therefore estimate the lower bound (9) as

$$\hat{D}(P) = D_{\text{EDSR}} + \left[(P_{\text{EDSR}} - P)_+\right]^2,$$

where D_{EDSR} is the MSE of EDSR, and P_{EDSR} is the estimated Gelbrich distance between EDSR reconstructions and ground-truth images. Note the unoccupied region under the estimated curve in Figure 2, which is indeed unattainable according to the theory.

We also present 11 estimators $\hat{X}_t$ which we construct by interpolation between EDSR and ESRGAN [29], $\hat{X}_t = tX_{\text{EDSR}} + (1-t)X_{\text{ESRGAN}}$. We observe (Figure 2) that estimators constructed using these two extreme points are closer to the optimal DP tradeoff than the evaluated methods. Also note that since ESRGAN does not attain 0-perception index, we are practically able to use negative values $t < 0$ to extrapolate better perception-quality estimators $\hat{X}_{-0.05}$ and $\hat{X}_{-0.1}$. In Figure 3 we present a visual comparison between SRGAN-$\text{VGG}_{2,2}$ [12] and our interpolated estimator $\hat{X}_{0.12}$. Both achieve roughly the same RMSE distortion (18.09 for SRGAN, 18.15 for $\hat{X}_{0.12}$), but our estimator achieves a lower perception index. Namely, by using interpolation, we manage to achieve improvement in perceptual quality, without degradation in distortion. The improvement in visual quality is also apparent in the figure. Additional visual comparisons can be found in the Appendix.

[2] All codes are freely available and provided by the authors. The BSD100 dataset is free to download for non-commercial research.

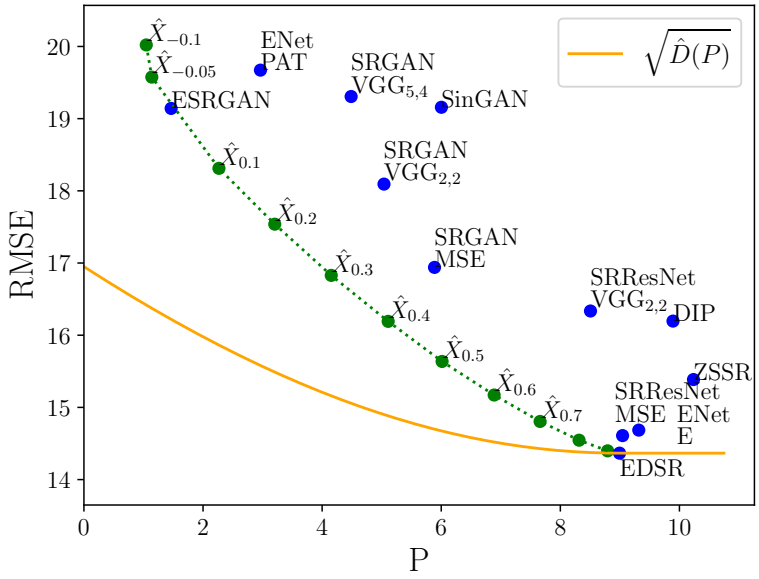

Figure 2: **Evaluation of SR algorithms**. We plot 12 algorithms (Blue) on the Distortion-Perception plane. Here we estimate perception using the Gelbrich distance between empirical means and covariances of the original data and reconstructed data. $\hat{D}(P)$ (Orange) is the estimated lower bound (9) where we consider EDSR to be the global minimizer X^*. Note the unoccupied region under the estimated curve, which is unattainable. We also plot 11 estimators $\hat{X}_t$ (Green) created by an interpolation between EDSR and ESRGAN, using different relative weights t. Note that estimators constructed using these two extreme estimators are closer to the optimal DP curve than the compared methods.

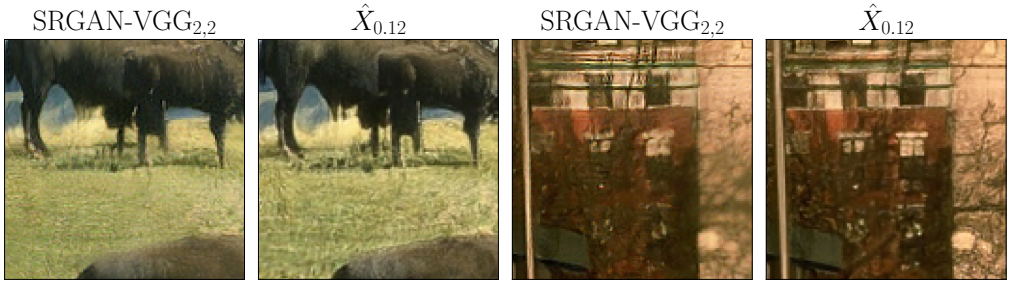

Figure 3: **A visual comparison between estimators with approximately the same MSE**. Left: SRGAN-VGG$_{2,2}$. Right: $\hat{X}_{0.12}$, an interpolation between EDSR and ESRGAN using $t = 0.12$. Observe the improvement in perceptual quality, without any significant degradation in distortion.

6 Conclusion

In this paper we provide a full characterization of the distortion-perception tradeoff for the MSE distortion and the Wasserstein-2 perception index. We show that optimal estimators are obtained by interpolation between the minimum MSE estimator and an optimal perfect perception quality estimator. In the Gaussian case, we explicitly formulate these estimators. To the best of our knowledge, this is the first work to derive such closed-form expressions. Our work paves the way towards fully understanding the DP tradeoff under more general distortions and perceptual criteria, and bridging between fidelity and visual quality at test-time, without training different models.

Broader impact This work deals with theoretical aspects of statistical estimation, and does not hold any foreseeable societal consequences.

References

[1] Mohamed Abderrahmen Abid, Ihsen Hedhli, and Christian Gagné. A generative model for hallucinating diverse versions of super resolution images. *arXiv preprint arXiv:2102.06624*, 2021.

[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] Yuval Bahat and Tomer Michaeli. Explorable super resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2716–2725, 2020.

[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] Xin Deng. Enhancing image quality via style transfer for single image super-resolution. *IEEE Signal Processing Letters*, 25(4):571–575, 2018.

[7] Roy Friedman and Yair Weiss. Posterior sampling for image restoration using explicit patch priors. *arXiv preprint arXiv:2104.09895*, 2021.

[8] Matthias Gelbrich. On a formula for the l2 wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

[9] Robert M Gray. Toeplitz and circulant matrices: A review. 2006.

[10] Bahjat Kawar, Gregory Vaksman, and Michael Elad. Stochastic image denoising by sampling from the posterior distribution. *arXiv preprint arXiv:2101.09552*, 2021.

[11] Martin Knott and Cyril S Smith. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43(1):39–49, 1984.

[12] 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.

[13] 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.

[14] Dong Liu, Haochen Zhang, and Zhiwei Xiong. On the classification-distortion-perception tradeoff. *arXiv preprint arXiv:1904.08816*, 2019.

[15] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Srflow: Learning the super-resolution space with normalizing flow. In *European Conference on Computer Vision*, pages 715–732. Springer, 2020.

[16] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423. IEEE, 2001.

[17] Pablo Navarrete Michelini, Dan Zhu, and Hanwen Liu. Multi–scale recursive and perception–distortion controllable image super–resolution. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[18] Guy Ohayon, Theo Adrai, Gregory Vaksman, Michael Elad, and Peyman Milanfar. High perceptual quality image denoising with a posterior sampling cgan. *arXiv preprint arXiv:2103.04192*, 2021.

[19] Victor M Panaretos and Yoav Zemel. *An invitation to statistics in Wasserstein space*. Springer Nature, 2020.

[20] Mangal Prakash, Alexander Krull, and Florian Jug. Divnoising: diversity denoising with fully convolutional variational autoencoders. *arXiv preprint arXiv:2006.06072*, 2020.

[21] Mangal Prakash, Mauricio Delbracio, Peyman Milanfar, and Florian Jug. Removing pixel noises and spatial artifacts with generative diversity denoising methods. *arXiv preprint arXiv:2104.01374*, 2021.

[22] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4491–4500, 2017.

[23] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4570–4580, 2019.

[24] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3118–3126, 2018.

[25] Alon Shoshan, Roey Mechrez, and Lihi Zelnik-Manor. Dynamic-net: Tuning the objective without re-training for synthesis tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3215–3223, 2019.

[26] Asuka Takatsu. On wasserstein geometry of gaussian measures. In *Probabilistic approach to geometry*, pages 463–472. Mathematical Society of Japan, 2010.

[27] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018.

[28] Michael Unser. On the approximation of the discrete karhunen-loeve transform for stationary processes. *Signal Processing*, 7(3):231–249, 1984.

[29] 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.

[30] Xintao Wang, Ke Yu, Chao Dong, Xiaoou Tang, and Chen Change Loy. Deep network interpolation for continuous imagery effect transition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1692–1701, 2019.

[31] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612, 2004.

[32] Zhen Zhang, Mianzhi Wang, and Arye Nehorai. Optimal transport in reproducing kernel hilbert spaces: Theory and applications. *IEEE transactions on pattern analysis and machine intelligence*, 42(7):1741–1754, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We mention it in Sec.1.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Our work consists of theoretical results.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] A detailed formulation of the problem and assumptions is given in Sec. 2. More specific assumptions are given in their relevant place.
   (b) Did you include complete proofs of all theoretical results? [Yes] Proofs for all theoretical results are given in the Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] The numerical results are presented only to illustrate the theory. They involve simple calculations and can be easily reproduced from the description in Sec. 5 (i.e. we didn't train any model).
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]