EXPLORING THE PARETO-OPTIMALITY BETWEEN QUALITY AND DIVERSITY IN TEXT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Quality and diversity are two essential aspects for performance evaluation of text generation models. Quality indicates how likely the generated samples are to be real samples, and diversity indicates how much differences there are between generated samples. Though quality and diversity metrics have been widely used for evaluation, it is still not clear what the relationship is between them. In this paper, we give theoretical analysis of a multi-objective programming problem where quality and diversity are both expected to be maximized. We prove that there exists a family of Pareto-optimal solutions, giving an explanation of the widely observed tradeoff behavior between quality and diversity in practice. We also give the structure of such solutions, and show that a linear combination of quality and diversity is sufficient to measure the divergence between the generated distribution and the real distribution. Further, we derive an efficient algorithm to reach the Pareto-optimal solutions in practice, enabling a controllable quality-diversity tradeoff.

1 INTRODUCTION

Text generation is an essential task for many NLP applications, such as machine writing (Zhang et al., 2017), machine translation (Bahdanau et al., 2014), image captioning (Rennie et al., 2017) and dialogue system (Li et al., 2017). Recently, lots of neural generation models have been proposed and gained increasing attentions (Yu et al., 2017; Fedus et al., 2018; Chen et al., 2018). However, it is still an open problem which metrics are suitable to evaluate the performance of text generation models. Among the metrics used in practice, generation quality and diversity are two most widely considered aspects. High generation quality requires the model to generate realistic samples, i.e. generated samples are free of grammatical or logical errors. While high generation diversity requires the model to generate diverse samples, i.e. generated samples are less likely to be duplicate and contain diverse unique patterns.

This work is motivated by three questions about quality and diversity:

- Q1: *What is the relationship between quality and diversity?* Besides being evaluation metrics, high generation quality and diversity have also been critical requirements in many applications (Li et al., 2015; Xu et al., 2018; Zhang et al., 2018b). However, many researches find that quality and diversity show a tradeoff behavior among well-trained models (Lu et al., 2018; Gao et al., 2019; Hashimoto et al., 2019). Though in accordance with intuition, such observations stay empirical and lack of theoretical support.
- Q2: *Is there any gap between quality-diversity evaluation and the divergence objective?* The original objective of training a text generation model is to approximate the probability distribution of real text data, which is equivalent to minimizing a divergence between model distribution and the real distribution (Mikolov et al., 2010). Since divergence would be intractable if the text probabilities are not modeled explicitly, some researchers opt to use quality and diversity metrics instead as a remedy (Fedus et al., 2018; Chen et al., 2018). However, it is not clear whether it is sufficient to approximate divergence by integration of quality and diversity.
- Q3: *How to achieve optimal solutions in practice when quality and diversity are both required to be maximized?* Quality or diversity may be focused more than another in some applica-

tions. Though researchers have proposed different methods to tackle different application scenarios (Zhang et al., 2018a; Li et al., 2015; Zhang et al., 2018b), it is still an open problem how to maximize one aspect while keeping another above some threshold, i.e. achieving a Pareto-optimal solution.

In this paper, we try to answer the above three question under the unconditional text generation setting. We first give a general definition of quality and diversity, and then study a Multi-Objective Programming(MOP) problem which maximizes quality and diversity simultaneously. Answers are given by performing theoretical analysis over this MOP problem:

A1: *Quality and diversity truly act as a tradeoff.* We prove there exists a family of Pareto-optimal solutions for the MOP problem, which constitutes the Pareto-frontier. For each Pareto-optimal solution Q, there exists another Pareto-optimal solution Q', such that either quality or diversity of Q' is higher than Q. This indicates that a quality-diversity tradeoff exists among all these optimal solutions, and non-optimal solutions have the potential to be improved over both metrics.

A2: *Quality and diversity can be combined to be a divergence.* We prove that a linear combination of some paired quality and diversity constitutes a divergence metric between the generated distribution and the real distribution, including some widely recognized quality and diversity metrics as special cases.

A3: *Optimal solutions over both quality and diversity can be achieved by our proposed QDTC method.* We prove that the optimal solutions of the MOP problem can be obtained by optimizing a designed objective function, and propose a QDTC algorithm which can be implemented efficiently like the widely used maximum likelihood estimation method. Experiments show that this algorithm achieves controllable quality and diversity tradeoff on both synthetic data and real MSCOCO dataset.

2 Related Work

To evaluate the performance of text generation models, many evaluation metrics are designed for different purposes. Early neural text generation models use Perplexity(PPL) to show how much a language model fit the training data (Mikolov et al., 2010), and this metric is still adopted in recent works (Zhang et al., 2018a; Fedus et al., 2018). PPL is correspondent to the Kullback-Leibler(KL) divergence, thus is a metric showing the difference between model distribution and the real distribution. However, Chen et al. (1998) show that PPL does not seem to correlate well with the task performance in real applications. Moreover, PPL cannot be calculated if text probabilities are not explicitly given. Therefore, the quality and diversity of generated text are further considered as complementary metrics.

For quality metrics, the evaluation is closely related to the ground truth distribution. Yu et al. (2017) propose to use Negative Log-Likelihood where the real distribution is known in advance, which measures the average log-probability of generated samples over the real distribution. If the real distribution is not explicitly given, BLEU (Papineni et al., 2002) and ROUGE (Lin & Och, 2004) are usually applied, which measure the n-gram overlap between generated samples and a set of reference ground truth samples. For diversity metrics, the evaluation is performed within the model itself. Li et al. (2015) proposed Distinct-n as diversity metric, which calculates the ratio of unique n-grams in generated samples. Zhu et al. (2018) proposed another metric called Self-BLEU, which is similar to BLEU score but use generated samples as reference set. In this work, we assume the real distribution P and the model distribution Q are explicitly given. To perform theoretical analysis, we propose a general form of quality and diversity, which is in accordance with above proposed metrics.

Although the tradeoff behavior between quality and diversity has not been well studied theoretically, there have been some works trying to control such tradeoff. The temperature-based method is the most widely used one (Hashimoto et al., 2019; Fan et al., 2018; Lau et al., 2017). By dividing the probability vector by a temperature factor t before softmax operation, one can achieve higher quality with smaller t and higher diversity with larger t during the decoding stage. Another method to control the tradeoff during training is proposed by Li et al. (2019). With different hyper-parameters in the objective function, the trained model can get higher quality at the expense of lower diversity.

Whether these methods can achieve optimal solutions under quality and diversity is still not clear, and the conclusions will be discussed in this paper.

3 DEFINITION OF QUALITY AND DIVERSITY

Currently there is no unified definition for quality and diversity in text generation, which poses great challenges for further theoretical studies. In fact, it is not easy to define a general form of quality and diversity due to various understandings of these two aspects. In this paper, we try to give a general form of quality and diversity in a mathematical view, though it may not be comprehensive enough to cover all possible understandings.

3.1 A GENERAL FORM OF QUALITY AND DIVERSITY

Text data is usually discrete, so we make the following notations. Assume the vocabulary size is $|V|$, and the maximum length is L, then the distribution of text data can be described by a categorical distribution with size $N = |V|^L$. We denote the real distribution and the generated model distribution as $P(x) = (P_1, P_2, \cdots, P_N)$ and $Q(x) = (Q_1, Q_2, \cdots, Q_N)$, respectively.

In general, the *Quality* of a text generation model measures how likely the generated text are to be realistic text in human's view. Since the value of real probability $P(x)$ can be viewed as reflecting the realistic degree of a text x, the expectation of some function over $P(x)$ could be used to quantify quality. For example, in Yu et al. (2017) and Nie et al. (2018), the *Log-Likelihood*(LL) is used as the quality metric, where $LL(Q; P) = \mathbb{E}_{x\sim Q} \log P(x)$. Following this idea, we propose a general form of quality, i.e., $U(Q; P) = \mathbb{E}_{x\sim Q} f_u[P(x)]$, where $f_u[P(x)]$ is a function over $P(x)$.

Similarly, the *Diversity* of a text generation model measures how much difference there are among those generated texts. From the viewpoint of information, *Shannon-Entropy(SE)* of $Q(x)$ can be used as a natural diversity metric, where $SE(Q) = -\mathbb{E}_{x\sim Q} \log Q(x)$. From another understanding view, a text x should be less likely to be generated again if the diversity is high. This idea has been adopted in biology to evaluate the diversity of biocoenosis, named as the *Simpson's Diversity Index(SDI)*, where $SDI(Q) = 1 - \mathbb{E}_{x\sim Q} Q(x)$. Summarizing these two different understandings, we obtain a general form of diversity, i.e. $V(Q) = -\mathbb{E}_{x\sim Q} f_v[Q(x)]$.

To this end, we propose a general form of quality and diversity metrics as follows:

$$U(Q) = U(Q; P) = \mathbb{E}_{x\sim Q} f_u[P(x)] = \sum_{i=1}^{N} Q_i \cdot f(P_i), \quad V(Q) = -\mathbb{E}_{x\sim Q} f_v[Q(x)] = \sum_{i=1}^{N} g(Q_i),$$

where $f_u(x)$ is denoted as $f(x)$ and $-\frac{f_v(x)}{x}$ is denoted as $g(x)$.

3.2 THE RATIONALITY OF QUALITY AND DIVERSITY

To guarantee U and V are rational quality and diversity metrics, we need to discuss about the conditions of f and g. Without loss of generality, we first assume that f is differentiable and g is twice differentiable. Further, the following requirements are necessary for rational quality and diversity:

1. Generating more samples with higher real probability yields higher overall quality;
2. Distributing the probability more equally yields higher overall diversity.

Mathematically, these two requirements can be formalized as the following two properties:

1. If $P_i > P_j$, then for $Q' = (Q_1, \ldots, Q_i + \epsilon, \ldots, Q_j - \epsilon, \ldots)$, $U(Q') > U(Q)$ for any $\epsilon \in (0, Q_j)$.

2. If $Q_i \geq Q_j$, then for $Q' = (Q_1, \ldots, Q_i + \epsilon, \ldots, Q_j - \epsilon, \ldots)$, $V(Q') < V(Q)$ for any $\epsilon \in (0, Q_j)$.

Then we can obtain the conditions of f and g by the following theorem:

Theorem 1. *The following conditions are both sufficient and necessary to satisfy the properties 1-2: For any x_1, x_2 s.t. $x_1 > x_2 > 0$ and $x_1 + x_2 \leq 1$, we have $f(x_1) > f(x_2)$ and $g'(x_1) < g'(x_2)$.*

According to Theorem 1, it is necessary for $f(x)$ to be strictly monotonically increasing and $g(x)$ to be strictly concave for $x \in (0, \frac{1}{2})$. For simplicity, we only consider the cases where such properties

hold for $x \in (0, 1)$, thus get a sufficient condition: i.e. $f(x)$ is strictly monotonically increasing for $x \in (0, 1)$, and $g(x)$ is strictly concave for $x \in (0, 1)$.

Under this condition, we can see that a model with highest quality will distribute all its density to text with highest real probability, and a model with highest diversity will be uniform, which are consistent with human understandings.

We list some speical cases under this condition, which will be used as examples in the following analysis. For quality metrics, we use *Log-Likelihood(LL)* with $f(x) = \log x$ and *Coverage-Rate(CR)* with $f(x) = x$. For diversity metrics, we use *Shannon-Entropy(SE)* with $g(x) = -x \log x$ and *Negative Repeat-Rate(NRR)* with $g(x) = -x^2$.

4 THE PARETO-OPTIMALITY

4.1 THE MOP PROBLEM

To explore the relationship between quality and diversity, we consider the following Multi-Objective Programming(MOP):

$$\max_{Q} (U(Q), V(Q))$$
$$s.t. \sum_{i=1}^{N} Q_i = 1$$
$$\forall i, \ -Q_i \leq 0$$

The goal is to maximize both quality and diversity, while keeping Q a legal distribution. The optimal solutions of a MOP problem are called Pareto-optima, which means no other solution can beat them consistently over all objectives.

We give definitions of the terminologies of Pareto-optimality below:

Definition 1. *For two distributions Q and Q', if one of the following conditions are satisfied, we say that Q is dominated by Q'.*

1. *$U(Q') > U(Q)$ and $V(Q') \geq V(Q)$;*
2. *$U(Q') \geq U(Q)$ and $V(Q') > V(Q)$.*

A solution Q is called a Pareto-optimum if it is not dominated by any Q'. The set containing all the Pareto-optima is called the Pareto-frontier.

Intuitively, a Pareto-optimum is a solution that there is no distribution can achieve both higher quality and higher diversity than it. And all the Pareto-optima constitutes the Pareto-frontier. The Pareto-frontier may collapse into one solution which leads to a global optimum, e.g. if P is uniform, the unique optimal solution would be $Q^* = P$. However it is often the case where the objectives in MOP problem cannot reach their optima consistently, thus there exists a family of optimal solutions. To verify the tradeoff behavior between quality and diversity, we need to prove the existence of such a family of optimal solutions, thus the structure of the Pareto-frontier under a non-uniform P is what we care about.

4.2 THE PARETO-FRONTIER

We try to show what the Pareto-optima look like by giving the following theorems:

Lemma 1. *If Q is a Pareto-optimum, the following conditions are satisfied: if $P_i > P_j$, then $Q_i \geq Q_j$; if $P_i = P_j$, then $Q_i = Q_j$.*

Theorem 2. *For a distribution Q, if P is not uniform, then:*

(1) The following condition is both sufficient and necessary for Q to be a Pareto-optimum: there exist real value $w \leq 0$ and b that for any $i = 1, \ldots, N$, there is

$$Q_i = \hat{g}'^{-1}[w \cdot f(P_i) + b], \tag{1}$$

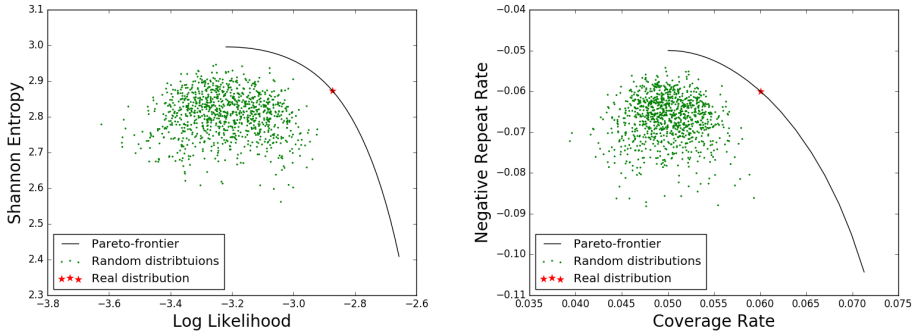

Figure 1: Illustration of the Pareto-frontier on a random toy categorical distribution with size 20. **Left:** The LL-SE case. **Right:** The CR-NRR case.

where

$$\hat{g}'^{-1}(x) = \begin{cases} g'^{-1}(x) & \text{if } x < g'(0), \\ 0 & \text{if } x \geq g'(0), \end{cases}$$

(2) b is correspondent to w, i.e. b is fixed once w is fixed. If $f(x) < 0$ for all $x \in [0, 1]$, then b is strictly monotonically increasing w.r.t. w. If $f(x) > 0$ for all $x \in [0, 1]$, then b is strictly monotonically decreasing w.r.t. w.

(3) If we denote a Pareto-optimum Q as $Q(w)$, then for any $w_1 < w_2$: if $w_1, w_2 \in [B, 0]$, there is $Q(w_1) \neq Q(w_2)$ and $U(Q(w_1)) > U(Q(w_2)), V(Q(w_1)) < V(Q(w_2))$; if $w_1, w_2 \in (-\infty, B]$, there is $Q(w_1) = Q(w_2)$; where $B = \frac{g'(\frac{1}{M}) - g'(0)}{f(P_{m_1}) - f(P_{m_2})}$, and $P_{m_1} = \max_i P_i$, $P_{m_2} = \max_{P_i \neq P_{m_1}} P_i$, $M = \#\{i | P_i = P_{m_1}\}$, # denotes the size of a set.

Lemma 1 shows that the optimal distribution is order-preserving, and Theorem 2 further gives the structure of Pareto-optima. Since different ws lead to different distributions, we can change w from 0 to B and get a family of optimal solutions with different quality and diversity. As such, for a non-uniform P, the Pareto-frontier is a family of distributions.

Now we can see that, if we want to maximize quality and diversity at the same time, these two metric acts as a tradeoff. Since all distributions in the Pareto-frontier are Pareto-optima, trying to improve one metric for an optimum will lead to another optimum at most, thus inevitably causing another metric to drop.

We show the result of Theorem 2 here on the special cases used in Section 3.2. We pair LL with SE, and CR with NRR. For the LL-SE metrics, the Pareto-optima can be written as

$$Q_i = \frac{P_i^\beta}{Z}, \; Z = \sum_{i=1}^{N} P_i^\beta, \; \beta \geq 0,$$

we have $w = -\beta$, and $b = 1 + \log Z$. This is exactly the case used in Li et al. (2019). For the CR-NRR metrics, the Pareto-optimum can be written as

$$Q_i = \frac{\max(P_i + \gamma, 0)}{Z}, \; Z = \sum_{i=1}^{N} \max(P_i + \gamma, 0), \; \gamma > -\max_i P_i,$$

we have $w = -\frac{2}{Z}$, and $b = -\frac{2\gamma}{Z}$. An illustration of the Pareto-frontier on a toy dataset is shown in Figure 1.

4.3 Relationship with Divergence

Besides quality and diversity, the direct difference between model distribution Q and real P is also considered in practice, which is usually evaluated by *Divergence* metrics such as the Kullback-Leibler divergence. Since calculation of divergence is usually intractable, quality and diversity are

often used together as a remedy. However, it is still not clear whether combining quality and diversity is sufficient for divergence evaluation.

We show that a linear combination of quality and diversity constitute a divergence metric if function f and g are carefully chosen. Define a weighted sum of quality and diversity as $W(Q) = \alpha U(Q) + (1-\alpha)V(Q), \alpha \in [0, 1)$, then $D(P||Q) = W(P) - W(Q)$ would become a divergence metric as long as $Q = P$ is a Pareto-optimum, as shown in the following Theorem:

Theorem 3. *The following condition is both sufficient and necessary for $Q = P$ to be in the Pareto-frontier for any P: there exist $w_0 \leq 0$ and b_0 that*

$$g(x) = w_0 \int f(x)\mathrm{d}x + b_0 x. \tag{2}$$

If the above condition is satisfied, then $Q = P$ corresponds to a Pareto-optimum with $w = w_0$ and $b = b_0$, and it is the only distribution that maximize $W(Q) = \alpha U(Q) + (1-\alpha)V(Q)$ with $\alpha = \frac{w_0}{w_0 - 1}$, and $D(P||Q) = W(P) - W(Q)$ becomes a divergence metric.

We find that if quality and diversity metrics are carefully chosen, namely g is the integral of a affine transformation of f, we can get a divergence metric by a linear combination of these two metrics. Since such condition is also necessary, the real distribution is unlikely to be a Pareto-optima if we use casually chosen metrics. This means, there would be one distribution achieving both higher quality and higher diversity than the ground truth, which is implausible. Illustration of such phenomenon with mismatched metrics is shown in Appendix A.7. Therefore, if the condition in Theorem 3 is not satisfied, it would be unlikely to measure the divergence using a combination of quality and diversity.

The special cases listed in Section 3.2 would satisfy the condition in Theorem 3 if LL is paired with SE and CR with NRR. For the LL-SE metrics, $D(P||Q) = \frac{1}{2}\sum_{i=1}^{N} Q_i \cdot \log \frac{Q_i}{P_i}$, which is exactly the Reverse KL divergence if the constant $\frac{1}{2}$ is ignored. For the CR-NRR metrics, $D(P||Q) = \frac{1}{3}\sum_{i=1}^{N}(Q_i - P_i)^2$, which measures the sum of squared difference among all probabilities.

5 Optimization of the MOP Problem

Though the original goal is to recover real distribution for text generation models, higher quality or higher diversity may become the primary requirement in real applications. As a result, it is meaningful to achieve other Pareto-optima besides recovering the real distribution, leading to a controllable quality-diversity tradeoff.

One widely used method for quality-diversity tradeoff control is introducing a temperature factor to the decoding stage of neural decoders. However, such temperature-based method violates the order-preserving requirements in Lemma 1, thus cannot achieve general Pareto-optima due to the sequential nature of text(see Appendix A.8 for explanation). In fact, it is non-trivial to achieve general Pareto-optima through such post-editing methods, i.e. train a model with $Q = P$ and then modify the decoding strategy.

As a result, we seek methods which can get the Pareto-optimal model immediately after training. In real applications, the real probability $P(x)$ is never explicitly given, thus a unified objective such as $W(Q)$ in Section 3.2 is not feasible for training a model. So we will give a method to achieve the Pareto-optima without knowing P.

5.1 Training Objective

Borrowing the idea from the DDR methodLi et al. (2019), we also use a modified training objective while keeping the algorithm similar to the widely used maximum likelihood estimation method. For a Pareto-optimum Q satisfying $Q_i = \hat{g}'^{-1}[w \cdot \hat{f}(P_i) + b]$, the corresponding training objective is

$$\begin{aligned} &\max_{Q} \mathbb{E}_{x\sim P}\, h[Q(x)], \\ &h(x) = \int \frac{c}{\hat{f}^{-1}[\frac{g'(x)-b}{w}]}\mathrm{d}x, \quad c > 0. \end{aligned} \tag{3}$$

Since f^{-1} has no definition outside of $[f(0), f(1)]$, we use $\hat{f}^{-1}$ as an expansion, and the value outside of $[f(0), f(1)]$ can be defined arbitrarily as long as $\hat{f}^{-1}$ is monotonically increasing and strictly positive. Theorem 4 gives the condition when such $\hat{f}^{-1}$ can be constructed and guarantees that we can get a Pareto-optimum by solving the above problem. In the following discussions, we further assume that f and g satisfy the conditions in Theorem 3 in order to get reasonable quality-diversity metrics.

Theorem 4. *There exists $\hat{f}^{-1}$ to make h concave, if and only if $\lim_{x\to 0^+} f(x) = -\infty$. If $h(x)$ is concave w.r.t $x \in (0, 1)$, then with a objective defined as Equation 3, the optimal solution is a Pareto-optimum defined as Equation 1.*

The parameter w and b in the expression of Pareto-optima provides a smooth way to control the quality-diversity tradeoff according to Theorem 2. Therefore, we can achieve higher quality or diversity by tuning w or b accordingly in Equation 3.

Since we do not know the value of both w and b for most of the time, such training objective cannot be applied directly to general cases. However, there are some cases which is still tractable. We observed that there is a free parameter c in the expression of $h(x)$. Since changing c does not change the solution, so if w or b could be separated from $h(x)$ and constitute a factor, we can get a feasible objective using another parameter. For example, if $h(x, w, b, c) = h_1(x, w) \cdot h_2(w, b) \cdot c$, we can set $c = h_2^{-1}(w, b)$ so that $h(x, w, b, c) = h_1(x, w)$. In this way, b can be neglected and we only need to care about w. According to Theorem 5, if f is the logarithmic function or the power function, then b or w can be neglected respectively.

Theorem 5. *If $\hat{f}^{-1} = f^{-1}$, then the following condition is both sufficient and necessary for $h(x, w, b, c)$ to be decomposed as $h(x, w, b, c) = h_1(x, w) \cdot h_2(w, b) \cdot c$: there exist constant a and d such that $f(x) = a \cdot \log x + d$.*

Also, the following condition is both sufficient and necessary for $h(x, w, b, c)$ to be decomposed as $h(x, w, b, c) = h_3(x, b) \cdot h_4(w, b) \cdot c$: there exist constant a and d such that $f(x) = d \cdot x^a$.

Theorem 5 provides a necessary condition for h to be used in practice, but we still need f to be monotonically increasing and h to be concave for a sufficient condition. Since an affine transformation of f is equivalent to f in terms of optimal solutions, we only consider the non-trivial cases in Theorem 5, including $f(x) = \log x$ and $f(x) = x^a$. To simplify the conclusion, we assume $g'(x) = -f(x)$ which means $w_0 = -1$ and $b_0 = 0$ in Theorem 3.

For the case of logarithmic function $f(x) = \log x$, we have $h'(x) = x^{\frac{1}{w}} \cdot e^{\frac{b}{w}} \cdot c$. And for the case of power function $f(x) = x^a$ where $a > 0$, we have $h'(x) = (x^a + b)^{-\frac{1}{a}} \cdot (-w)^{\frac{1}{a}} \cdot c$. This case does not satisfy the concavity condition, thus should be discarded. However, the continuity holds for $f(x) = -x^a$ where $a < 0$, we have $h'(x) = (x^a - b)^{-\frac{1}{a}} \cdot (-w)^{\frac{1}{a}} \cdot c$.

As such, we can select an appropriate c to diminish the factor w or b. Thus in practice we can use

$$h'(x) = x^{\frac{1}{w}},$$

or

$$h'(x) = (x^a - b)^{-\frac{1}{a}}, a < 0.$$

The derivative is sufficient for the gradient calculation, so it is not necessary to know the exact form of $h(x)$.

5.2 ALGORITHM

We show how to do the optimization using the expression of h'. For a model Q_θ parameterized by θ, denote the loss function as

$$\mathcal{L} = -\mathbb{E}_{x\sim P}\, h[Q_\theta(x)].$$

The gradient w.r.t θ at current value $\theta = \theta_0$ would be

$$\begin{aligned}
\nabla_\theta \mathcal{L}|_{\theta=\theta_0} &= -\mathbb{E}_{x\sim P}\, h'[Q_\theta(x)] \cdot \nabla_\theta Q_\theta(x)|_{\theta=\theta_0} \\
&= -\mathbb{E}_{x\sim P}\, h'[Q_\theta(x)] \cdot Q_\theta(x) \cdot \nabla_\theta \log Q_\theta(x)|_{\theta=\theta_0} \\
&= -\nabla_\theta \mathbb{E}_{x\sim P}\, h'[Q_{\theta_0}(x)] \cdot Q_{\theta_0}(x) \cdot \log Q_\theta(x)|_{\theta=\theta_0}.
\end{aligned}$$

Algorithm 1 The Quality-Diversity Tradeoff Control Algorithm

Input: Dataset $\mathcal{D} = \{x_i\}_{i=1}^N$, batch size M, learning rate α, model Q_θ, function h'.
1: Initialize Q_θ with random weights.
2: Pre-train Q_θ with Maximum Likelihood Estimation. (optional)
3: **repeat**
4: Sample M examples $\{x_i\}_{i=1}^M$ from $\mathcal{D}$.
5: $\theta_0 \leftarrow \theta$.
6: Calculate $T(x_i)$ for each i using Equation 4.
7: $\theta \leftarrow \theta + \alpha \cdot \nabla_\theta \frac{1}{M} \sum_{i=1}^M T(x_i) \cdot \log Q_\theta(x_i)$
8: **until** convergence

Table 1: A summary of the two QDTC methods used in our experiments.

Method	$f(x)$	$g(x)$	$h'(x)$	$U(Q)$	$V(Q)$
QDTC-logarithm	$\log x$	$-x \log x$	$x^{\frac{1}{w}}$	$\sum_{i=1}^N Q_i \log P_i$	$-\sum_{i=1}^N Q_i \log Q_i$
QDTC-reciprocal	$-\frac{1}{x}$	$\log x$	$\frac{1}{x} - b$	$-\sum_{i=1}^N \frac{Q_i}{P_i}$	$\sum_{i=1}^N \log Q_i$

Let

$$T(x) = h'[Q_{\theta_0}(x)] \cdot Q_{\theta_0}(x), \tag{4}$$

then

$$\nabla_\theta \mathcal{L}|_{\theta=\theta_0} = -\nabla_\theta \mathbb{E}_{x\sim P}\, T(x) \cdot \log Q_\theta(x)|_{\theta=\theta_0}. \tag{5}$$

Now the model can be optimized using Equation 5. We summarized this Quality-Diversity Tradeoff Control(QDTC) algorithm as Algorithm 1:

Note that when we use $h'(x) = x^{\frac{1}{w}}$ and constrain w in $(-\infty, -1)$, QDTC method would be equivalent to the Differentiated Distribution Recovery(DDR) method used by Li et al. (2019), thus DDR is a special case of our QDTC.

6 Experiments

In this section, we evaluate our proposed QDTC method on synthetic data as well as MSCOCO Image Caption dataset(Chen et al., 2015), compared with the temperature-based method.

For the temperature-based method, we pre-train the model with Maximum Likelihood Estimation(MLE), and then tune the temperature t for different output during decoding.

For our QDTC method, we use two pairs of metrics: the logarithm ones where $f(x) = \log x, g(x) = -x \log x$; and the reciprocal ones where $f(x) = -\frac{1}{x}, g(x) = \log x$. We summarize the details of these two cases in Table 1. Although QDTC can be used without any pre-training, we find it would converge faster and more stably if we pre-train the model with MLE. As a result, we also use MLE pre-training for QDTC in all of our experiments.

6.1 Experiments on Synthetic Data

In the synthetic data, the real probability P is explicitly given, so we can evaluate how close a generated model Q is to the Pareto-frontier.

Specifically, we define a sequential data space with vocabulary size 10 and length 3. Thus the total number of feasible texts is $N = 10^3$. The synthetic data are generated using a randomly initialized oracle model, whose parameters are known in advance. In our experiments, this model contains an embedding layer with dimension 32, an LSTM layer with 32 hidden nodes, and a fully-connected(FC) output layer with 10 hidden nodes.

The text generation model share the same structure with the oracle model, but use learned parameters. To guarantee the consistency between training and test, we do not construct a dataset $\mathcal{D}$ in advance. Instead, we sample data from the oracle model directly whenever data are needed. The

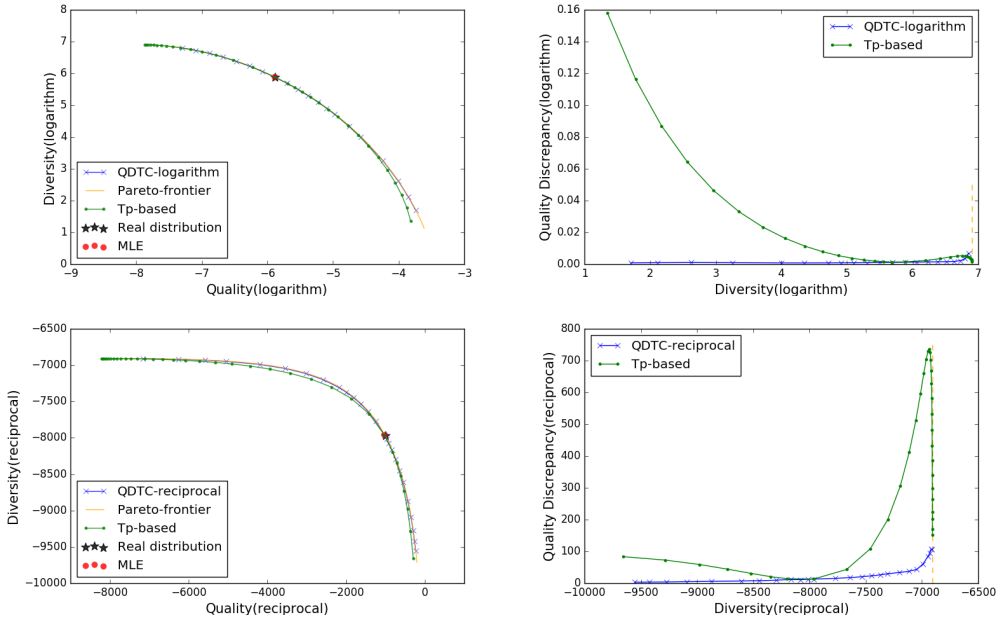

Figure 2: Evaluation of quality and diversity on synthetic data. Vertical dashed lines show the boundary of maximum diversity. **Left:** The original metrics. **Right:** The quality discrepancy under the same diversity level compared with the Pareto-frontier.

quality and diversity of the trained model are computed by the corresponding metrics in the logarithm case and the reciprocal case as shown in Table 1.

As we can see from the results shown in Figure 2, all methods show smooth curves from upper left to lower right under the evaluation of quality and diversity, indicating a tradeoff relation between the two metrics. The curves of our QDTC methods closely fit their corresponding ground truth curve, which means Pareto-optima are well obtained. The curve of temperature-based method gets close with ground truth at two points: the middle point where $t = 1$ and the leftmost point where $t \to \infty$. This makes sense because temperature-based method can achieve $Q = P$ with $t = 1$ and Q becomes uniform when t tends to $+\infty$. However at other points, the discrepancies grow much larger, indicating a failure to achieve other Pareto-optima.

6.2 Experiments on MSCOCO Dataset

To show the effectiveness of QDTC on real text data, we run experiments on the MSCOCO Image Caption dataset. Our empirical settings are exactly the same as Guo et al. (2017), including the preprocessing and the data separation. Specifically, only the captions are used as text data, and sentences which contain words with frequency lower than 10 are removed. 80,000 unique sentences are sampled as training set, and another 5,000 unique sentences are used as test set. The final vocabulary size is 4,840 and maximum text length is 32.

The architectures of the text generation models are similar as that on synthetic data. The embedding dimension and number of LSTM hidden nodes are set to 128, and the number of FC hidden nodes is 4,840.

Since the ground truth distribution P is unknown under this setting, the calculations of our general defined quality and diversity metrics may become intractable. Fortunately, the CR-NRR metrics can be approximated by sampling due to the linearity of f:

$$CR(Q;P) = \sum_{i=1}^{N} Q_i \cdot P_i = \mathbb{E}_{x\sim P}\, Q(x), \quad NRR(Q) = -\sum_{i=1}^{N} Q_i^2 = -\mathbb{E}_{x\sim Q}\, Q(x).$$

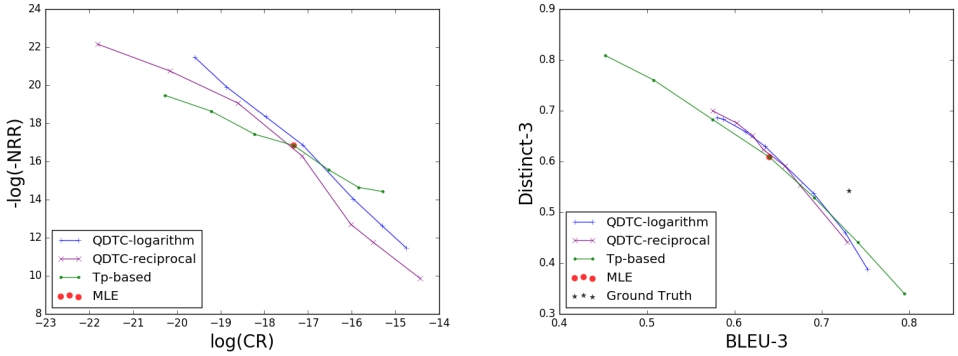

Figure 3: Evaluation of quality and diversity on MSCOCO dataset. We apply 7 hyper-parameters for each method, their corresponding values from left to right are: $[1.3, 1.2, 1.1, 1.0, 0.9, 0.8, 0.7]$ for t in temperature-based method; $[-0.85, -0.9, -0.95, -1.0, -1.1, -1.2, -1.3]$ for w in QDTC-logarithm; $[1e9, 1e8, 1e7, 0, -1e6, -2e6, -5e6]$ for b in QDTC-reciprocal.

Therefore, the expectation over Q in NRR can be directly taken on generated samples, while the expectation over P in CR can be calculated by sampling from the test set.

Besides using CR and NRR as metrics, we also evaluate our results by the widely used quality and diversity metrics in application, i.e. BLEU-n (Papineni et al., 2002) and Distinct-n (Li et al., 2015). BLEU-n measures the degree of n-gram overlap between generated text and a reference text set, i.e. the test set. Distinct-n calculates the ratio of unique n-grams over all n-grams in generated text. Here we set $n = 3$. The experimental results are shown in Figure 3.

From the results, we can see that with the change of w or b, our QDTC methods show smooth control of the quality-diversity tradeoff under CR-NRR and even BLEU-Distinct metrics. Therefore, QDTC can be applied in some real applications where quality or diversity is preferred while keeping another metric above a threshold. QDTC methods do not show consistent superiority over temperature-based method in the figure, this is because the metrics used in the real data are different from the corresponding theoretical metrics in QDTC. Nevertheless, QDTC performs better than temperature-based method in many cases.

7 Conclusion and Discussion

In this paper, we mainly focus on the theoretical study of quality and diversity in text generation. We give a general definition of quality and diversity, and then study the MOP problem where quality and diversity are both required to be maximized. Three main conclusions are obtained by our study:

Firstly, quality and diversity show a clear tradeoff relation in theory. Therefore, we suggest using both metrics for evaluation in real application instead of focusing only one metric, to get a comprehensive understanding of a specific text generation model.

Secondly, a linear combination of some paired quality and diversity is equivalent to a divergence. This theoretical result indicates that quality and diversity metrics should be carefully chosen in practice, to avoid the mistake that ground-truth distribution is non-optimal.

Thirdly, an algorithm named QDTC is proposed to efficiently optimize both quality and diversity. Experimental results show that QDTC achieves good approximation of the Pareto-optima on both synthetic data and real MSCOCO data. In applications where one metric is favored more than another, a good model should be able to achieve the required Pareto-optimal solution. We can see that our proposed QDTC gives a feasible example of how to achieve a controllable quality-diversity tradeoff in this direction.

In the future, we would like to study the relationship between quality and diversity under the conditional text generation settings. It is also anticipated to extend the conclusions to continuous data generation settings, such as image or video generation.

REFERENCES

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Liqun Chen, Shuyang Dai, Chenyang Tao, Haichao Zhang, Zhe Gan, Dinghan Shen, Yizhe Zhang, Guoyin Wang, Ruiyi Zhang, and Lawrence Carin. Adversarial text generation via feature-mover's distance. In *Advances in Neural Information Processing Systems*, pp. 4666–4677, 2018.

Stanley F Chen, Douglas Beeferman, and Ronald Rosenfeld. Evaluation metrics for language models. In *DARPA Broadcast News Transcription and Understanding Workshop*, pp. 275–280. Citeseer, 1998.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.

William Fedus, Ian Goodfellow, and Andrew M Dai. Maskgan: Better text generation via filling in the _. *arXiv preprint arXiv:1801.07736*, 2018.

Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. Jointly optimizing diversity and relevance in neural response generation. *arXiv preprint arXiv:1902.11205*, 2019.

Jiaxian Guo, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. Long text generation via adversarial training with leaked information. *arXiv preprint arXiv:1709.08624*, 2017.

Tatsunori B Hashimoto, Hugh Zhang, and Percy Liang. Unifying human and statistical evaluation for natural language generation. *arXiv preprint arXiv:1904.02792*, 2019.

Jey Han Lau, Timothy Baldwin, and Trevor Cohn. Topically driven neural language model. *arXiv preprint arXiv:1704.08012*, 2017.

Jianing Li, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. Differentiated distribution recovery for neural text generation. In *AAAI*, 2019.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*, 2017.

Chin-Yew Lin and FJ Och. Looking for a few good metrics: Rouge and its evaluation. In *Ntcir Workshop*, 2004.

Sidi Lu, Yaoming Zhu, Weinan Zhang, Jun Wang, and Yong Yu. Neural text generation: Past, present and beyond. *arXiv preprint arXiv:1803.07133*, 2018.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

Weili Nie, Nina Narodytska, and Ankit Patel. Relgan: Relational generative adversarial networks for text generation. 2018.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Association for Computational Linguistics, 2002.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, volume 1, pp. 3, 2017.

Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu Sun. Diversity-promoting gan: A cross-entropy based generative adversarial network for diversified text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3940–3949, 2018.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, pp. 2852–2858, 2017.

Hainan Zhang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. Reinforcing coherence for sequence to sequence model in dialogue generation. In *IJCAI*, pp. 4567–4573, 2018a.

Hainan Zhang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. Tailored sequence to sequence models to different conversation scenarios. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1479–1488, 2018b.

Jiyuan Zhang, Yang Feng, Dong Wang, Yang Wang, Andrew Abel, Shiyue Zhang, and Andi Zhang. Flexible and creative chinese poetry generation using neural memory. *arXiv preprint arXiv:1705.03773*, 2017.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 1097–1100. ACM, 2018.

A Appendix

A.1 Preliminaries

Before starting the proofs, we first introduce some preliminaries on the constrained convex optimization problem. Assume $f(x)$, $c_i(x)$, and $h_j(x)$ are continuous differentiable function define on $\mathbb{R}^n$, consider the constrained convex optimization problem defined as follows:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f(x) \\ s.t. \quad & c_i(x) \leq 0, \quad i = 1, 2, \cdots, k \\ & h_j(x) = 0, \quad j = 1, 2, \cdots, l \end{aligned} \tag{6}$$

The optimal solutions for above problem are given by, as shown in the following theorem:

Theorem 6. *Assume $f(x)$ and $c_i(x)$ are convex, $h_j(x)$ are affine, and c_i are strictly feasible (there exists one x satisfying $c_i(x) < 0$ for all i). Define the Lagrange function as:*

$$L(x, \alpha, \beta) = f(x) + \sum_{i=1}^{k} \alpha_i c_i(x) + \sum_{j=1}^{l} \beta_j h_j(x), \tag{7}$$

where $\alpha \geq 0$. Then the the following conditions are both sufficient and necessary for x to be a solution in problem 6.

$$\begin{aligned} \nabla_x L(x^*, \alpha^*, \beta^*) &= 0 \\ \nabla_\alpha L(x^*, \alpha^*, \beta^*) &= 0 \\ \nabla_\beta L(x^*, \alpha^*, \beta^*) &= 0 \\ \alpha_i^* c_i(x^*) &= 0, \quad i = 1, 2, \cdots, k \\ c_i(x^*) &\leq 0, \quad i = 1, 2, \cdots, k \\ \alpha_i^* &\geq 0, \quad i = 1, 2, \cdots, k \\ h_j(x^*) &= 0, \quad j = 1, 2, \cdots, k \end{aligned} \tag{8}$$

The conditions in Equation 8 are called the Karush-Kuhn-Tucker(KKT) conditions.

A.2 PROOF OF THEOREM 1

For property 1, from $U(Q') - U(Q) = \epsilon f(P_i) - \epsilon f(P_j) = \epsilon[f(P_i) - f(P_j)] > 0$, we get $f(P_i) > f(P_j)$. We then get the conclusion by setting $x_1 = P_i$ and $x_2 = P_j$.

For property 2, $V(Q') - V(Q) = [g(Q_i + \epsilon) + g(Q_j - \epsilon)] - [g(Q_i) + g(Q_j)] < 0$ is true for any $Q_i > Q_j$. Denote $C = Q_i + Q_j$ and $r(x) = g(x) + g(C - x)$, then we have $V(Q') - V(Q) = r(Q_i + \epsilon) - r(Q_i) < 0$ for any Q_i, ϵ. Since $0 < Q_i < Q_i + \epsilon < 1$, we need $r'(x) < 0$ for $x \in (0, 1)$. Then, since $r'(x) = g'(x) - g'(C - x) < 0$ is true for any $0 < C - x < x < 1$. Set $x_1 = C - x$ and $x_2 = x$ and we get $g'(x_1) < g'(x_2)$ for any $x_1 > x_2 > 0$ and $x_1 + x_2 = Q_i + Q_j \leq 1$.

A.3 PROOF OF LEMMA 1

If $P_i > P_j$, assume $Q_i < Q_j$, we can construct Q' where $Q'_k = Q_k$ for all $k \neq i, j$ and $Q'_i = Q_j, Q'_j = Q_i$. As such, $V(Q') = V(Q)$ but $U(Q') - U(Q) = (Q_j - Q_i)[f(P_i) - f(P_j)] > 0$. This means Q is dominated by Q', which conflicts with the fact that Q is a Pareto-optimum. So $Q_i \geq Q_j$.

If $P_i = P_j$, assume $Q_i \neq Q_j$, and we can further assume $Q_i > Q_j$. Again we construct Q' where $Q'_k = Q_k$ for all $k \neq i, j$ and $Q'_i = Q'_j = \frac{Q_i + Q_j}{2}$. Surely we have $U(Q') = U(Q)$, and $V(Q') - V(Q) = 2g(\frac{Q_i + Q_j}{2}) - g(Q_i) - g(Q_j)$. Since g is strictly concave, we have $V(Q') - V(Q) > 0$, which means Q is dominated by Q'. This causes confliction, so $Q_i = Q_j$.

A.4 LEMMA 2 AND ITS PROOF

This lemma is used to support the proof of Theorem 2 and Theorem 3.

Lemma 2. *Assume $\alpha \in [0, 1)$ and $W(Q) = \alpha U(Q) + (1 - \alpha)V(Q)$, then the distribution Q that maximize $W(Q)$ satisfies $Q_i = \hat{g}'^{-1}[w \cdot f(P_i) + b]$, and $w = \frac{\alpha}{\alpha - 1}$.*

Define the optimization problem as follows:

$$\min_Q -\alpha \cdot U(Q) - (1 - \alpha)V(Q)$$
$$s.t.\ 1 - \sum_{i=1}^{N} Q_i = 0$$
$$\forall i,\ -Q_i \leq 0$$

Again we first check that the prerequisites in KKT are all satisfied. $-U(Q)$ is linear and $-V(Q)$ is convex w.r.t. Q'; $1 - \sum_{i=1}^{N} Q_i$ is affine w.r.t. Q'; since all Q_i can be positive, so the inequalities are all strictly feasible.

The Lagrange function is:

$$L(Q_i, \lambda, \xi_i) = -\alpha \sum_{i=1}^{N} Q_i f(P_i) - (1 - \alpha) \sum_{i=1}^{N} g(Q_i) + \lambda(1 - \sum_{i=1}^{N} Q_i) - \sum_{i=1}^{N} \xi_i Q'_i, \quad \xi \geq 0$$

Apply KKT and we get the following conditions for a optimal solution:

$$\forall i,\ \frac{\partial L}{\partial Q_i} = -\alpha f(P_i) - (1 - \alpha)g'(Q_i) - \lambda - \xi_i = 0,$$
$$\forall i,\ -\xi_i Q_i = 0$$

For $Q_i \neq 0$, there is $\xi_i = 0$, so

$$Q_i = g'^{-1}[\frac{\alpha}{\alpha - 1} f(P_i) + \frac{\lambda}{\alpha - 1}];$$

for $Q_i = 0$, there is $\xi_i > 0$, so

$$\frac{\alpha}{\alpha - 1} f(P_i) + \frac{\lambda}{\alpha - 1} > g'(0).$$

Denote $w = \frac{\alpha}{\alpha-1}$ and $b = \frac{\lambda}{\alpha-1}$ and combine the two cases together, we get:

$$Q_i = \hat{g}'^{-1}[w \cdot f(P_i) + b], \quad w \leq 0,$$

The above derivation is both sufficient and necessary, so we finished the proof.

A.5 Proof of Theorem 2

We give the proofs for three conclusions individually.

A.5.1 Conclusion 1

Here we only consider the case with $U(Q) \neq \max_Q U(Q)$, and the case where $U(Q) = \max_Q U(Q)$ will be incorporated into conclusion 3. We try to find a distribution Q' with the highest diversity while quality is not lower than Q. Define a convex optimization problem as follows:

$$\begin{aligned} \min_{Q'} & -V(Q') \\ s.t.\ & U(Q) - U(Q') \leq 0 \\ & 1 - \sum_{i=1}^{N} Q'_i = 0 \\ & \forall i,\ -Q'_i \leq 0 \end{aligned}$$

For Q to be a Pareto-optimum, it's both sufficient and necessary for Q to be a solution of above problem. Thus we try to solve this problem next.

We first check that the prerequisites in KKT are all satisfied. $-V(Q')$ is convex w.r.t. Q'; $1 - \sum_{i=1}^{N} Q'_i$ is affine w.r.t. Q'; $U(Q) - U(Q')$ and $-Q'_i$ are convex(linear) w.r.t Q'; since all Q'_i can be positive and $U(Q) \neq \max_Q U(Q)$, so the inequalities are all strictly feasible.

The Lagrange function is:

$$L(Q_i, \lambda, \eta, \xi_i) = -\sum_{i=1}^{N} g(Q'_i) + \lambda(1 - \sum_{i=1}^{N} Q'_i) + \eta \sum_{i=1}^{N} (Q_i - Q'_i) f(P_i) - \sum_{i=1}^{N} \xi_i Q'_i, \quad \eta, \xi \geq 0$$

Apply KKT and we get the following conditions for a optimal solution:

$$\begin{aligned} \forall i,\ \frac{\partial L}{\partial Q'_i} = -g'(Q'_i) - \lambda - \eta f(P_i) - \xi_i = 0, \\ \eta[U(Q) - U(Q')] = 0, \\ \forall i,\ -\xi_i Q'_i = 0 \end{aligned}$$

Since we need Q to be a solution, so

$$\begin{aligned} \forall i,\ -g'(Q_i) - \lambda - \eta f(P_i) - \xi_i = 0, \\ \forall i,\ -\xi_i Q_i = 0 \end{aligned}$$

For $Q_i \neq 0$, there is $\xi_i = 0$, so $Q_i = g'^{-1}[-\eta f(P_i) - \lambda]$; for $Q_i = 0$, there is $\xi_i > 0$, so $-\eta f(P_i) - \lambda > g'(0)$. Denote $w = -\eta$ and $b = -\lambda$ and combine the two cases together, we get:

$$Q_i = \hat{g}'^{-1}[w \cdot f(P_i) + b], \quad w \leq 0,$$

where

$$\hat{g}'^{-1}(x) = \begin{cases} g'^{-1}(x) & \text{if } x < g'(0), \\ 0 & \text{if } x \geq g'(0). \end{cases}$$

The above derivations are both sufficient and necessary, thus we finish the proof.

A.5.2 CONCLUSION 2

For the second conclusion, we separate the proof into two parts: (1) b is correspondent to w; (2) the monotonicity of b w.r.t. w.

(1) The sum of all Q_i should be 1. Denote

$$T(w,b) = \sum_{i=1}^{N} \hat{g}'^{-1}[w \cdot f(P_i) + b].$$

Since $g'(x)$ is strictly monotonically decreasing, so $T(w,b)$ is monotonically non-increasing w.r.t. b. If $T(w,b) > 0$, there would be a term which is strictly monotonically decreasing w.r.t. b, under which condition $T(w,b)$ is strictly monotonically decreasing w.r.t. b. Also, $T(w,b)$ is continuous w.r.t. b since g'^{-1} is continuous. When

$$b = g'(0) - w \cdot f(\max_i P_i),$$

there is

$$w \cdot f(P_i) + b \geq w \cdot f(\max_i P_i) + b = g'(0),$$

so $T(w,b) = 0$; when

$$b = g'(\frac{1}{N}) - w \cdot f(\min_i P_i),$$

there is

$$w \cdot f(P_i) + b \leq w \cdot f(\min_i P_i) + b = g'(\frac{1}{N}),$$

so $T(w,b) \geq 1$. From above analysis, the value of T can reach 0 or be greater than 1. So combining the monotonicity of T, there exists and only one b that satisfies $T(w,b) = 1$, leading to a rational distribution.

(2) Define $T(w,b) = \sum_{i=1}^{N} \hat{g}'^{-1}[w \cdot f(P_i) + b(w)]$ as above. Since $T(w,b)$ represents the total probability of a distribution, so there should be $T(w,b) \equiv 1$, thus $\frac{\mathrm{d}T}{\mathrm{d}w} = 0$.

$$\frac{\mathrm{d}T}{\mathrm{d}w} = \sum_{i \in S} \frac{f(P_i) + b'(w)}{g''\{g'^{-1}[w \cdot f(P_i) + b(w)]\}},$$

where $S = \{i | w \cdot f(P_i) + b(w) < g'(0)\}$. By the condition $\frac{\mathrm{d}T}{\mathrm{d}w} = 0$, we get

$$b'(w) = -\frac{\sum_{i \in S} \frac{f(P_i)}{g''\{g'^{-1}[w \cdot f(P_i) + b(w)]\}}}{\sum_{i \in S} \frac{1}{g''\{g'^{-1}[w \cdot f(P_i) + b(w)]\}}}.$$

Since $g''(x) < 0$, so if $f(x) < 0$ for all $x \in [0,1]$, we can get $b'(w) > 0$, thus b is strictly monotonically increasing w.r.t. w. Similarly, if $f(x) > 0$ for all $x \in [0,1]$, we can get $b'(w) < 0$, thus b is strictly monotonically decreasing w.r.t. w.

A.5.3 CONCLUSION 3

For the third conclusion, we also separate the proof into two parts: (1) the uniqueness of $Q(w)$; (2) the monotonicity of U and V w.r.t. w.

(1) Since P is not uniform, so we can denote B, P_{m_1}, P_{m_2} as they are in the theorem. According to Lemma 1, since P_{m_1} is the largest one, so the corresponding Q_{m_1} is also the largest one, which means

$$Q_{m_1} = \hat{g}'^{-1}[w \cdot f(P_{m_1}) + b] > 0.$$

Thus we get

$$w \cdot f(P_{m_1}) + b < g'(0).$$

At the same time, because we can get $Q_i = Q_{m_1}$ if $P_i = P_{m_1}$, so we can sum up all the largest Q_i and get

$$M \cdot Q_{m_1} \leq \sum_{i=1}^{N} Q_i = 1,$$

we can get

$$w \cdot f(P_{m_1}) + b \geq g'(\frac{1}{M}). \tag{9}$$

Consider the case where $w \geq B$, we first prove that $w \cdot f(P_{m_2}) + b \leq g'(0)$. Assume

$$w \cdot f(P_{m_2}) + b > g'(0), \tag{10}$$

then $Q_{m_2} = 0$, and there is $Q_i = 0$ for any i satisfying $P_i \leq P_{m_2}$. As a result, there should be $Q_i = \frac{1}{M}$ for all i satisfying $P_i = P_{m_1}$, which means

$$w \cdot f(P_{m_1}) + b = g'(\frac{1}{M}). \tag{11}$$

Subtract Equation 11 by Equation 10, we get

$$w \cdot [f(P_{m_1}) - f(P_{m_2})] < g'(\frac{1}{M}) - g'(0),$$

so

$$w < \frac{g'(\frac{1}{M}) - g'(0)}{f(P_{m_1}) - f(P_{m_2})} = B.$$

This contradict with the fact that $w \geq B$. Thus we have $w \cdot f(P_{m_2}) + b \leq g'(0)$.

Combining the above conclusions, for any $w_1, w_2 \in [B, 0]$, assume $Q(w_1) = Q(w_2)$, then

$$w_1 \cdot f(P_{m_1}) + b_1 = w_2 \cdot f(P_{m_1}) + b_2,$$
$$w_1 \cdot f(P_{m_2}) + b_1 = w_2 \cdot f(P_{m_2}) + b_2.$$

As $P_{m_1} \neq P_{m_2}$, so $w_1 = w_2$, causing contradiction. Thus we have $Q(w_1) \neq Q(w_2)$.

For any $w \leq B$, assume

$$w \cdot f(P_{m_2}) + b < g'(0). \tag{12}$$

By subtracting Equation 9 and Equation 12, we get

$$w \cdot [f(P_{m_1}) - f(P_{m_2})] > g'(\frac{1}{M}) - g'(0),$$

so

$$w > \frac{g'(\frac{1}{M}) - g'(0)}{f(P_{m_1}) - f(P_{m_2})} = B.$$

This causes contradiction, so the above assumption does not hold. Thus we have $w \cdot f(P_{m_2}) + b \geq g'(0)$, which means $Q_{m_2} = 0$. Borrowing the proof above, we know that $Q_i = \frac{1}{M}$ for all i satisfying $P_i = P_{m_1}$. This is a trivial Pareto-optimal case where $U(Q) = \max_Q U(Q)$. Now we know the distribution Q is fixed and does not change as w changes, so for any $w_1, w_2 \leq B$, there is $Q(w_1) = Q(w_2)$.

(2) For the expression of Q_i, since f and g' are both continuous and monotonic, so it is easy to know that Q_i is continuous w.r.t. w, then $U(Q(w))$ and $V(Q(w))$ are both continuous w.r.t. w. We just need to prove the monotonicity.

Assume $B \leq w_1 < w_2 \leq 0$, the goal is to prove that $U(Q(w_1)) > U(Q(w_2))$ and $V(Q(w_1)) < V(Q(w_2))$. According to Lemma 2, w_1 and w_2 have their corresponding $\alpha_1 = \frac{w_1}{w_1 - 1}$ and $\alpha_2 = \frac{w_2}{w_2 - 1}$, and $\alpha_1 > \alpha_2$. Since $Q(w)$ is the optimal solution for problem $\alpha U(Q) + (1 - \alpha)V(Q)$, and $Q(w_1)$ is different with $Q(w_2)$, so the following inequalities hold:

$$\alpha_1 U(Q(w_1)) + (1 - \alpha_1)V(Q(w_1)) > \alpha_1 U(Q(w_2)) + (1 - \alpha_1)V(Q(w_2)),$$
$$\alpha_2 U(Q(w_1)) + (1 - \alpha_2)V(Q(w_1)) < \alpha_2 U(Q(w_2)) + (1 - \alpha_2)V(Q(w_2)).$$

Subtracting the first equation by the second one, we get

$$(\alpha_1 - \alpha_2)[(U(Q(w_1)) - U(Q(w_2))) - (V(Q(w_1)) - V(Q(w_2)))] > 0.$$

As $\alpha_1 > \alpha_2$, so

$$U(Q(w_1)) - U(Q(w_2)) > V(Q(w_1)) - V(Q(w_2)).$$

Because $Q(w_1)$ and $Q(w_2)$ are both Pareto-optima, there quality and diversity should satisfy one of the following: $U(Q(w_1)) > U(Q(w_2)), V(Q(w_1)) < V(Q(w_2))$ or $U(Q(w_1)) < U(Q(w_2)), V(Q(w_1)) > V(Q(w_2))$. With the derived restriction $U(Q(w_1)) - U(Q(w_2)) > V(Q(w_1)) - V(Q(w_2))$, we know the first one holds, that is $U(Q(w_1)) > U(Q(w_2))$ and $V(Q(w_1)) < V(Q(w_2))$.

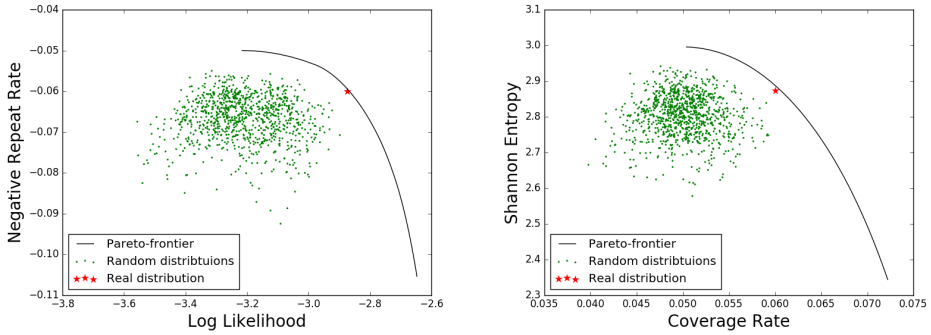

Figure 4: Illustration of the Pareto-frontier on a random toy categorical distribution with size 20. The diversity metrics are swaped, thus are mismatched. **Left:** Pair LL with NRR. **Right:** Pair CR with SE. Note that there is always a gap between the star and the curve, indicating that the real distribution lies on neither of the two Pareto-frontiers.

A.6 Proof of Theorem 3

The requirement that $Q = P$ in the Pareto-frontier is equivalent to the following condition: for any P, there exist $w_0 \leq 0$ and b_0 that for any i, there is

$$P_i = \hat{g'}^{-1}[w_0 \cdot f(P_i) + b_0].$$

This means, for any $P_i > 0$, there is $w_0 \cdot f(P_i) + b_0 = g'(P_i)$. Since f and g' are both continuous, so

$$w_0 \cdot f(0) + b_0 - g'(0) = \lim_{P_i \to 0} w_0 \cdot f(P_i) + b_0 - g'(P_i) = 0.$$

We can see $w_0 \cdot f(P_i) + b_0 = g'(P_i)$ is also true for $P_i = 0$. By solving this differential equation, we get

$$g(x) = w_0 \int f(x)\mathrm{d}x + b_0 x.$$

Here b_0 can be any value because $P_i = {g'}^{-1}[w_0 \cdot f(P_i) + b_0]$ always lead to a plausible distribution P. Under this condition, we know that $Q = P$ is the only distribution that maximize $W(Q) = \alpha U(Q) + (1-\alpha)V(Q)$ where $\alpha = \frac{w_0}{w_0 - 1}$ according to Lemma 2. With the above conclusions, it is easy to check that $D(P||Q) = W(P) - W(Q) \geq 0$ and $D(P||Q) = 0$ if and only if $Q = P$, thus $D(P||Q)$ is a divergence metric.

A.7 Illustration of Pareto-frontier with Mismatched Metrics

We show in Figure 4 that the point $Q = P$ is under the Pareto-frontier curve when quality and diversity metrics are not matched, i.e. the condition in Theorem 3 is not satisfied. We use the same toy dataset as in Figure 1, but pair LL with NRR and CR with SE.

A.8 Insufficiency of Temperature-based Method

Temperature-based method for quality-diversity tradeoff is implemented through dividing the probability vector $\boldsymbol{a}$ by a temperature factor t before the softmax operation:

$$Q(\boldsymbol{x}; t) = Softmax(\boldsymbol{a}/t) = \frac{1}{Z}(e^{a_1/t}, \cdots, e^{a_{|v|}/t}), \quad Z = \sum_{i=1}^{|V|} e^{a_i/t}.$$

When the temperature is high ($t \to \infty$), the distribution would be near uniform; and when the temperature is low ($t \to 0$), the distribution would be sharp and assign $Q(x) \to 1$ for x with largest $P(x)$, and $Q(x) \to 0$ for other x. However in text generation, this is applied for the conditional probability of each token given its prefix, i.e. $P(x_i|x_{1:i-1})$, and is intractable to apply on the global probability $P(x_{1:L})$.

We give an example showing that this method violates the requirements in Lemma 1, thus cannot be used to achieve general Pareto-optima. To be concrete, if the post-edited model distribution is Q, we construct a simple case that $Q_i > Q_j$ while $P_i < P_j$ for some i, j.

Assume the text length is 2, and the vocabulary is $\{a, b, c\}$. Then the real probability of a text sample $x = (x_1, x_2)$ would be $P(x) = P(x_1) \cdot P(x_2|x_1)$. We assume $P(x_1) = (0.5, 0.5, 0.0), P(x_2|x_1 = a) = (0.4, 0.3, 0.3), P(x_2|x_1 = b) = (0.5, 0.5, 0.0)$, then we have $P(aa) = 0.2 < P(ba) = 0.25$. However when the temperature t approaches 0, there would be $Q(x_1) = (0.5, 0.5, 0.0), Q(x_2|x_1 = a) = (1.0, 0.0, 0.0), Q(x_2|x_1 = b) = (0.5, 0.5, 0.0)$, so that $Q(aa) = 0.5 > Q(ba) = 0.25$.

This clearly violates the order-preserving nature of a Pareto-optimum, thus we can never get the full Pareto-frontier using the temperature-based method, no matter what f and g are.

A.9 Proof of Theorem 4

We discuss the condition of f as conclusion 1 and the optimal solutions of the objective as conclusion 2.

A.9.1 Conclusion 1

Since $g'(x)$ is continuous and strictly monotonically decreasing w.r.t. x and $w < 0$, $\frac{g'(x)-b}{w}$ is continuous and strictly monotonically increasing w.r.t. x. By changing the value of w and b, $\frac{g'(x)-b}{w}$ can reach any finite real value. To make h concave, we need $\frac{1}{\hat{f}^{-1}(x)}$ to be monotonically decreasing w.r.t. $x \in (-\infty, \infty)$.

We know that f is strictly monotonically increasing for $x \in (0, 1)$, so $f(0) = \lim_{x\to 0^+} f(x)$ is either a finite value or $-\infty$.

Consider the case where $f(0)$ is a finite value z. The domain of definition of f^{-1} would be $(z, f(1))$, and $f^{-1}(z) = 0$. Since $\hat{f}^{-1}$ in an expansion of f^{-1}, $\hat{f}^{-1}$ shares the same value in $(z, f(1))$, so $\hat{f}^{-1}(z) = 0$. This will lead to $\lim_{x\to z^+} \frac{1}{\hat{f}^{-1}(x)} = +\infty$. Thus $\frac{1}{\hat{f}^{-1}(x)}$ is unable to be monotonically decreasing near $x = z$.

Then we consider the case where $f(0) = -\infty$. The domain of definition of f^{-1} would be $(-\infty, f(1)]$, and $f^{-1}(x) > 0$ for any $x \in (-\infty, f(1)]$. If $f(1) = +\infty$, then we construct $\hat{f}^{-1}(x) = f^{-1}(x)$, so that $\hat{f}^{-1}(x)$ is strictly positive and monotonically increasing. As such, $\frac{1}{\hat{f}^{-1}(x)}$ would be monotonically decreasing. If $f(1)$ is a finite value, which means $\hat{f}^{-1}(f(1)) = 1$, it is easy to define the value of $\hat{f}^{-1}(x)$ in $x \in (f(1), +\infty)$ to guarantee $\hat{f}^{-1}(x)$ is monotonically increasing and strictly positive. As such, $\frac{1}{\hat{f}^{-1}(x)}$ would also be monotonically decreasing and strictly positive.

Based on above analysis, we can conclude that it is both sufficient and necessary for $\lim_{x\to 0^+} f(x) = -\infty$.

A.9.2 Conclusion 2

This problem is a convex optimization problem:

$$\begin{aligned} \min_{Q} & -\sum_{i=1}^{N} P_i \cdot h(Q_i) \\ s.t.\ & 1 - \sum_{i=1}^{N} Q_i = 0 \\ & \forall i,\ -Q_i \le 0 \end{aligned}$$

The Lagrange function is:

$$L(Q, \lambda, \xi_i) = -\sum_{i=1}^{N} P_i h(Q_i) + \lambda(1 - \sum_{i=1}^{N} Q_i) - \sum_{i=1}^{N} \xi_i Q_i, \quad \xi \ge 0.$$

We then check if $Q_i = \hat{g}'^{-1}[w \cdot f(P_i) + b]$ is a solution of this problem. Since $h'(x) = \frac{c}{\hat{f}^{-1}[\frac{g'(x)-b}{w}]}$, so

$$\frac{\partial L}{\partial Q_i} = -\frac{c \cdot P_i}{\hat{f}^{-1}[\frac{g'(Q_i)-b}{w}]} - \lambda - \xi_i.$$

We show that this requirement is satisfied with $Q_i = \hat{g}'^{-1}[w \cdot f(P_i) + b]$, $\lambda = -c$: if $Q_i > 0$, then

$$g'(Q_i) = w \cdot f(P_i) + b,$$

in which case $\xi_i = 0$, and

$$\frac{\partial L}{\partial Q_i} = -\frac{c \cdot P_i}{P_i} - \lambda - \xi_i = 0.$$

If $Q_i = 0$, then

$$w \cdot f(P_i) + b \geq g'(0),$$

so

$$0 \leq \frac{P_i}{\hat{f}^{-1}[\frac{g'(0)-b}{w}]} \leq 1.$$

We can set

$$\xi_i = -\frac{c \cdot P_i}{\hat{f}^{-1}[\frac{g'(0)-b}{w}]} + c \geq 0,$$

so that

$$\frac{\partial L}{\partial Q_i} = 0.$$

The above derivations are both sufficient and necessary, so $Q_i = \hat{g}'^{-1}[w \cdot f(P_i) + b]$ is the unique optimal solution.

A.10 Proof of Theorem 5

We discuss the logarithm function case as conclusion 1 and the power function case as conclusion 2.

A.10.1 Conclusion 1

For sufficiency, when $f(x) = a \log x + d$, there is $f^{-1}(x) = \exp\{\frac{x-d}{a}\}$, so

$$\begin{aligned} h'(x, w, b, c) &= \frac{c}{f^{-1}[\frac{g'(x)-b}{w}]} \\ &= \frac{c}{\exp\{\frac{g'(x)}{wa} - \frac{b+wd}{wa}\}} \\ &= \exp\{-\frac{g'(x) - wd}{wa}\} \cdot \exp\{\frac{b}{wa}\} \cdot c. \end{aligned}$$

By setting $h_1(x, w) = \int \exp\{-\frac{g'(x)-wd}{wa}\}\mathrm{d}x$ and $h_2(w, b) = \exp\{\frac{b}{wa}\}$, we get $h(x, w, b, c) = \int h'(x, w, b, c)\mathrm{d}x = h_1(x, w) \cdot h_2(w, b) \cdot c$.

For necessity, pick any $x_1, x_2 \in (0, 1)$, and construct a function:

$$H(x, w, b) = \frac{h'(x_1, w, b, c)}{h'(x_2, w, b, c)} = \frac{f^{-1}[\frac{g'(x_2)-b}{w}]}{f^{-1}[\frac{g'(x_1)-b}{w}]}.$$

If the decomposition can be found, this will require $H(x, w, b)$ to exclude b in its expression, so there should be

$$\frac{\partial H(x, w, b)}{\partial b} = 0.$$

Denote $A(x, w, b) = f^{-1}[\frac{g'(x)-b}{w}]$. Then we have

$$\frac{\partial H(x, w, b)}{\partial b} = \frac{1}{A(x_2, w, b)^2} \cdot \{-\frac{A(x_1, w, b)}{wf'[A(x_2, w, b)]} + \frac{A(x_2, w, b)}{wf'[A(x_1, w, b)]}\} = 0.$$

This lead to

$$A(x_1, w, b) \cdot f'[A(x_1, w, b)] = A(x_2, w, b) \cdot f'[A(x_2, w, b)].$$

This holds for any x_1, x_2, so both the left and right side should be a constant. Denote this constant as a, then

$$A(x, w, b) \cdot f'[A(x, w, b)] = a.$$

For simplicity, denote $A(x, w, b)$ as t, then

$$t \cdot f'(t) = a.$$

We get $f(x) = a \log x + d$ by solving this differentiable equation, where a, d can be any constant.

A.10.2 CONCLUSION 2

For sufficiency, when $f(x) = d \cdot x^a$, there is $f^{-1}(x) = (\frac{x}{d})^{\frac{1}{a}}$, so

$$\begin{aligned} h'(x, w, b, c) &= \frac{c}{f^{-1}[\frac{g'(x)-b}{w}]} \\ &= \frac{c}{[\frac{g'(x)-b}{wd}]^{\frac{1}{a}}} \\ &= [g'(x) - b]^{-\frac{1}{a}} \cdot (wd)^{\frac{1}{a}} \cdot c. \end{aligned}$$

By setting $h_1(x, b) = \int [g'(x) - b]^{-\frac{1}{a}} \mathrm{d}x$ and $h_2(w, b) = (wd)^{\frac{1}{a}}$, we get $h(x, w, b, c) = \int h'(x, w, b, c) \mathrm{d}x = h_1(x, b) \cdot h_2(w, b) \cdot c$.

For necessity, we again pick any $x_1, x_2 \in (0, 1)$, and construct the function:

$$H(x, w, b) = \frac{h'(x_1, w, b, c)}{h'(x_2, w, b, c)} = \frac{f^{-1}[\frac{g'(x_2)-b}{w}]}{f^{-1}[\frac{g'(x_1)-b}{w}]}.$$

If the decomposition can be found, this will require $H(x, w, b)$ to exclude w in its expression, so there should be

$$\frac{\partial H(x, w, b)}{\partial w} = 0.$$

Again denote $A(x, w, b) = f^{-1}[\frac{g'(x)-b}{w}]$. Then we have

$$\frac{\partial H(x, w, b)}{\partial b} = \frac{1}{A(x_2, w, b)^2} \cdot \{-\frac{A(x_1, w, b)[g'(x_2) - b]}{w^2 f'[A(x_2, w, b)]} + \frac{A(x_2, w, b)[g'(x_1) - b]}{w^2 f'[A(x_1, w, b)]}\} = 0.$$

This lead to

$$A(x_1, w, b) \cdot f'[A(x_1, w, b)] \cdot \frac{w}{g'(x_1) - b} = A(x_2, w, b) \cdot f'[A(x_2, w, b)] \cdot \frac{w}{g'(x_2) - b}.$$

This holds for any x_1, x_2, so both the left and right side should be a constant. Denote this constant as a, then

$$A(x, w, b) \cdot f'[A(x, w, b)] \cdot \frac{w}{g'(x) - b} = a.$$

For simplicity, denote $A(x, w, b)$ as t, then $\frac{g'(x)-b}{w} = f(t)$, and

$$t \cdot f'(t) \cdot \frac{1}{f(t)} = a.$$

We get $f(x) = d \cdot x^a$ by solving this differentiable equation, where a, d can be any constant.